"Regression, classification and feature selection from survival data : modeling of hypoxia conditions for cancer prognosis"

Branders, Samuel

Abstract

An important aspect of cancer research is the development of better prognostic tools for clinicians. These tools aim at predicting the survival time outcome of patients. Such tools are crucial as they assist clinicians in the choice of the best treatment strategy for each patient. An accurate prognostic model could help to save patients from unnecessary treatments. This thesis is centred around the development of prognostic models in their three key aspects. These three aspects are the selection of relevant markers, the learning of the prognosis model itself and its validation. We proposed the Coxlogit model for feature selection from survival and classification data. Classification and survival prediction are two common tasks in cancer research. With the Coxlogit model, we propose to model together these two tasks to improve the prediction and the feature selection. The Coxlogit model can be seen as a regularized mixture of a Cox and logistic models. The relevance of a prognostic mo...

Document type : *Thèse (Dissertation)*

Université catholique de Louvain
Institute of Information and Communication Technologies, Electronics and Applied Mathematics
Machine Learning Group

Regression, Classification and Feature Selection from Survival Data

—

Modeling of hypoxia conditions for cancer prognosis

Samuel Branders

Thesis presented for the Ph.D. degree
in Engineering Sciences

Thesis jury :
Prof. **Pierre Dupont** (Université catholique de Louvain), Advisor
Prof. **Olivier Feron** (Université catholique de Louvain), Advisor
Prof. **Jean-Baptiste Demoulin** (Université catholique de Louvain)
Prof. **Gianluca Bontempi** (Université Libre de Bruxelles)
Prof. **Sauerbrei Willi** (University of Freiburg)
Prof. **Pecheur Charles** (Université catholique de Louvain), President

September 2015

version of September 3, 2015

‘ Learning is not compulsory... neither is survival. ’

William Edwards Deming

Acknowledgements

En premier, je voudrais remercier mes promoteurs Pierre et Olivier. Sous ta supervision Pierre, j'ai énormément appris dont pas mal de soft skills qui vont m'être bien utile dans ma future carrière. J'ai adoré les nombreuses discussions que nous avons eu ensemble. Tu es à peu près aussi têtu que moi, ce qui les a rendu très intéressantes. :-)

Olivier, j'ai beaucoup apprécié travailler avec toi et ton équipe. Nos différences de "culture" ont nourri de riches discussions durant nos nombreux meetings et m'ont permis de découvrir plein de choses.

I also would like to thank all the members of my thesis commitee and jury for their feedback. All your questions and remarks helped me to improve this work.

Je voudrais remercier tous les collègues d'INGI, doctorants, académiques ou encore l'équipe PAT. En particulier, Jey, Adrien et Roberto, c'était un véritable plaisir de travailler avec vous.

Je remercie également tous les membres du machine learning group dont Michel, John, Guillaume, Benoît, Dimitri, Emilie et Alexandra. Je garde d'excellents souvenirs de nos ESANN, barbecues, etc.

Merci aussi à tous les potes avec qui j'ai pu partager le plaisir de faire une thèse (et bien plus): Antoine, Fwé, Gbb, Jb, Jey, Julien, Ka, Minou, Nico, Simon et Xa.

En dernier, je voudrais remercier toute ma famille. Papa, Maman et Papy, vous avez forgé ma passion pour les sciences. Je ne serais pas là sans vous. Nath, sans toi et ton soutien au quotidien, cette thèse aurait été beaucoup plus difficile.

Enfin, j'ai une pensée particulière pour ma marraine et mes proches qui ont souffert d'un cancer. Vous avez été une incroyable source de motivation. J'ai envie de croire que mon travail aurait pu vous être utile.

Abstract

An important aspect of cancer research is the development of better prognostic tools for clinicians. These tools aim at predicting the survival time outcome of patients. Such tools are crucial as they assist clinicians in the choice of the best treatment strategy for each patient. An accurate prognostic model could help to save patients from unnecessary treatments.

This thesis is centred around the development of prognostic models in their three key aspects. These three aspects are the selection of relevant markers, the learning of the prognosis model itself and its validation.

We proposed the Coxlogit model for feature selection from survival and classification data. Classification and survival prediction are two common tasks in cancer research. With the Coxlogit model, we propose to model together these two tasks to improve the prediction and the feature selection. The Coxlogit model can be seen as a regularized mixture of a Cox and logistic models.

The relevance of a prognostic model is typically assessed with a hazard ratio between the predicted risk groups. We identified some limitations of the hazard ratio in this particular context. More precisely, it appears to be very sensitive to the choice of discretization of the risk scores and has extreme values with unbalanced risk groups. We investigate the effect of the discretization in risk groups for the hazard ratio and other related metrics. A new metric, the balanced hazard ratio, is also proposed to solve those issues.

The biomedical part of this thesis investigates the use of hypoxia related gene signatures as potential prognostic markers. In controlled experiments, cell lines were submitted to normoxia, hypoxia and cycling hypoxia and then used to deduce molecular signatures. Promising prognosis results were found on real breast cancer data. Moreover, these hypoxia related gene signatures turn out to be an added value to the standard clinical prognostic models.

Contents

Chapter 1

Introduction

Cancer is a major cause of death in the world. In Europe only, the number of new cases was estimated to more than 3 millions for 2012 [46]. The most common ones are the female breast, the colorectal, prostate and lung cancer. An important aspect of cancer research is the development of better prognostic tools for clinicians [63, 106, 127, 64]. These tools aim at predicting the survival time outcome of patients [3], computing their risk to develop metastasis after surgery. Such tools are crucial as they assist clinicians in the choice of the best treatment strategy for each patient. For instance, one might decide according to the estimated risk whether a patient needs an adjuvant chemotherapy, a radiotherapy or both. An accurate prognostic model could help to save patients from unnecessary treatments. Treatments that have both high costs and adverse side effects.

The traditional prognostic models are based on clinicopathological criteria including, for example, the tumor size, histological grade, nodal status, age, etc. Usually, a small number of them are combined together to form a prognostic model. This combination in a model outputs a score representing the risk of each patient, the risk score. The higher is the risk score, the greater are the chances of relapse. An example of traditional prognostic models is the Nottingham prognostic index (NPI) for breast cancer [54]. The NPI model computes the risk as a linear combination of the tumor size, the histological grade and nodal status.

In the last 20 years, new technologies were developed, like the microarray and next-generation sequencing (NGS). In one experiment, they are able to measure the expression levels of thousands of genes. The expression level measure the transcription rate of a gene, which can be seen as its activity. Those technologies offer us a 'view' of the tumor at

a molecular level. This view gives precious information about the cancer biology and could help to improve the traditional prognostic tools. This improvement comes through the identification of new prognostic markers based on gene expression. Not limited any more to the small set of clinically measured variables, new prognostic models bring the hope of better and personalized treatment strategies.

This thesis is centred around the development of prognostic models in its three key aspects. These three aspects are the selection of relevant markers, the learning of the prognosis model itself and its validation. They all falls within the general framework of survival analysis. Survival analysis is thus introduce here and is followed by a brief description of the thesis contributions. These contributions concern both methodological aspects and biomedical findings. As a machine learning thesis more emphasis shall be put on the methodological part.

1.1 Survival analysis

Survival analysis is the class of statistical methods used to study the occurrence and timing of event. It is used in many domains like engineering, sociology, economics, etc. Some examples are the study of the time to industrial component failure, criminal recidivism, divorce or school graduation. In cancer research, survival analysis studies the time from the diagnosis or treatment to the occurrence of some event. This event of interest, also called the end-point, can be for instance the relapse, the death, or the appearance of distant metastasis.

A specific feature of such data is the censoring. A patient is censored if the particular event of interest has not been observed for him/her. Hopefully in cancer with good prognosis, most patients do not have metastasis and do not die before the end of the follow-up. Those patients are thus censored. The only available information for those patients is a lower bound of their survival time. Other reasons can explain such a censoring. Some patients move away or do not want to be in the study anymore. They could also die from other unrelated causes known as competing risks. This specificity of survival data makes the use of standard regression techniques inappropriate. Inappropriate in the sense that they can not make a proper use of censored patients.

There are two central functions used to describe survival data: the survivor function and the hazard function. These functions are so important that their standard estimation methods are the two most cited statistical papers of all time [140]. The survivor function models the probability of a patient to survive longer than some time t. This function

is the standard way to represent the survival of patients, the well-known survival curves. The hazard function, on the other hand, is used to express and model the risk of patients. The hazard function computes the risk of a patient to experience the event at some time t knowing that he has not experienced the event so far.

The Cox proportional hazard model is the most common way to model hazard functions. It allows us to compute the relationship between variables and the survival of patients through their hazard. The success of this method comes from the possibility to model the relative risk of patients without actually estimating full hazard functions. In practice, the Cox proportional hazard model is the method of choice for both the learning and the validation of prognostic models. The Cox model is thus central in the context of this thesis and will be covered in detail, in chapters 2 and 3.

1.2 Contributions

The methodological and biomedical contributions, summarized here, are described in detail respectively in Part II and Part III of this thesis. Often, the methodological contributions were triggered by problems and needs of real biomedical applications, especially in cancer research. They will be mostly presented in this context. However, those contributions aim at a broader applicability even outside the biomedical domains.

1. **Coxlogit model: feature selection from survival and classification data.** A common task in cancer research consists in showing and understanding the links between some patient conditions and their survival. One way is to find biomarkers that are predictive of both the patient condition and survival. The feature selection for the prediction of classes/conditions or survival are two common tasks in cancer research. However, they are generally tackled independently. The Coxlogit model is an embedded feature selection to select markers jointly predictive for both tasks seen as a multi-objective optimization. The Coxlogit model can be seen as a regularized mixture of a Cox and logistic models. The approach is detailed in the Chapter 6.

2. **Random non-linear projections for survival analysis.** The Cox proportional hazard model cannot account directly for non-linearity effects in patient variables. Several approaches relying on the use of kernels or neural networks have been proposed to tackle this issue. We proposed to use the cox model with non-linear

random projections as used in extreme learning. The simplicity and efficiency of this approach spares the additional complexity of defining an appropriate non-linear kernel or of training complex neural networks. Moreover the results of the model are still interpretable in terms of hazard. The random non-linear projections for survival analysis are presented in Chapter 7.

3. **Balanced hazard ratio for risk group evaluation from survival data.** The classification of patients in risk groups is needed since clinicians are routinely required to decide whether a specific treatment should be considered for a given patient. Such decision is precisely based on the assignment of this patient to a particular risk group. Besides, the definition of risk groups also comes from the stratification of patients according to discrete categories (e.g. smoker/non-smoker, gender, ...). The relevance of such risk groups is typically assessed by their difference in survival computed with a hazard ratio. We identified some limitations of the hazard ratio in this particular context. More precisely, it appears to be very sensitive to the choice of discretization of the risk scores and has extreme values with unbalanced risk groups. The hazard ratio can also be inconsistent with other metrics like the logrank test. In Chapter 8, we investigate the effect of the discretization in risk groups for the hazard ratio and other related metrics. A new metric, the balanced hazard ratio, is also proposed to solve those issues.

4. **Hypoxia signatures for cancer prognosis.** Hypoxia, the lack of oxygen, is a common characteristic of tumors. The tumor responses to this lack of O_2 are well known: tumor angiogenesis and glycolytic metabolism. These two mechanisms induce a local and temporal O_2 fluctuation, known as cycling hypoxia. The extent of cycling hypoxia reflects the tumor plasticity and capacity to survive and proliferate. Cycling hypoxia can thus be an important prognostic biomarker of cancer progression. However, technologies to measure these fluctuations are not easily available on real tissues for clinicians.

   Part III of this thesis investigates the use of hypoxia related gene signatures as potential prognostic markers. In controlled experiments, cell lines were submitted to normoxia, hypoxia and cycling hypoxia and then used to deduce molecular signatures. Promising prognosis results were found on real breast cancer data. Moreover, these hypoxia related gene signatures turn out to be an added value

to the standard clinical prognostic models.

Those contributions are the object of some publications and patent:

- Branders, S. and Dupont, P. (2015). A balanced hazard ratio for risk group evaluation from survival data. *Statistics in Medicine*, **34**(17)

- Branders, S., Frénay, B., and Dupont, P. (2015b). Survival Analysis with Cox Regression and Random Non-linear Projections. *Proceedings of the 23th European Symposium on Artificial Neural Networks*, pages 119–124

- Branders, S., D'Ambrosio, R., and Dupont, P. (2014). The Coxlogit model: Feature selection from survival and classification data. In *2014 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making (MCDM)*, pages 137–143

- Branders, S., D'Ambrosio, R., and Dupont, P. (2015a). A mixture Cox-Logistic model for feature selection from survival and classification data. *arXiv:1502.01493 [stat.ML]*, pages 1–6

- Boidot, R., Branders, S., Helleputte, T., Rubio, L. I., Dupont, P., and Feron, O. (2014). A generic cycling hypoxia-derived prognostic gene signature: application to breast cancer profiling. *Oncotarget*, **5**(16)

- Grandjean, M., Sermeus, A., Branders, S., Defresne, F., Dieu, M., Dupont, P., Raes, M., De Ridder, M., and Feron, O. (2013). Hypoxia Integration in the Serological Proteome Analysis Unmasks Tumor Antigens and Fosters the Identification of Anti-PhosphoeEF2 Antibodies as Potential Cancer Biomarkers. *PloS one*, **8**(10), e76508

- Seront, E., Rottey, S., Sautois, B., Kerger, J., D'Hondt, L. a., Verschaeve, V., Canon, J.-L., Dopchie, C., Vandenbulcke, J. M., Whenham, N., Goeminne, J. C., Clausse, M., Verhoeven, D., Glorieux, P., Branders, S., Dupont, P., Schoonjans, J., Feron, O., and Machiels, J.-P. (2012). Phase II study of everolimus in patients with locally advanced or metastatic transitional cell carcinoma of the urothelial tract: clinical activity, molecular response, and biomarkers. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, **23**(10), 2663—-2670

- Feron, O., Boidot, R., Branders, S., Dupont, P., and Helleputte, T. (2015). Signature of cycling hypoxia and use thereof for the prognosis of cancer. WO Patent App. PCT/EP2014/066,643

Part I

Background

Chapter 2

Survival Analysis

Survival analysis is the class of statistical methods used to study the occurrence and timing of events. It is used in many domains like engineering, sociology, economics and medicine. In cancer research, survival analysis studies the time from the diagnosis or treatment to the occurrence of some event [28]. This event of interest, also called the end-point, is usually relapse, cancer related death, or appearance of distant metastasis. For the sake of simplicity, the term survival is used here even with non-lethal end-points.

Survival analysis is used to answer many questions: What is the survival probability of patients at 10 years? What are the differences in survival between patients treated with different medications? What are the risk factors explaining early recurrences?

The medical context of this thesis allows us to make some assumptions on the survival data encountered. Those assumptions are very common and fit in the classical survival analysis framework. They naturally guide the choices of the statistical methods that are used and presented here.

Many possible kinds of events exist, but an event is always assumed to be a switch from one state to another at a precise location in time. Even if it can be relevant in some other situations, we assume here that an event can only appears once for each patient/sample.

In some study, multiple end-points are recorded such as death from cancer and death from heart attack. Such mutually exclusive events are known as competing risks. Specific survival techniques are designed to deal with such data. However, competing risks are rarely recorded in practice. We decide not to consider them in this chapter.

One may ask why there is a need for specific survival methods? The

time of event is a continuous variable and could thus be studied by classical regression techniques. The specificity of the survival data comes from the censoring. A patient is censored if the particular event of interest has not been observed for him/her. There are different kinds of censoring and many reasons to explain it. This chapter starts with more detailed explanations on censored data, in section 2.1.

The survival time of patients can be regarded as a random variable. A random variable which is commonly described using its probability density, survivor and hazard functions. Section 2.2 presents those key functions/distributions on which are based the different survival methods and models. Two of them, the Kaplan-Meier method and Cox proportional hazard model, are introduced respectively in section 2.3 and 2.4. More advanced survival regression techniques are detailed in the chapter 3.

2.1 Censored data

In survival data, there are many patients with no observed event. There are many possible reasons for such a situation. The patient may not experience the event before the end of the study. He could also die from another unrelated cause or simply have moved away. Those patients lost to follow-up are called censored patients.

Three kinds of censoring can be distinguished. The previous examples correspond to right censored patients. The only available information for those patients is a lower bound of their survival time.

Sometimes, a patient could experience the event before the first examination after the treatment/diagnosis. In such a case, the only available information is an upper bound on the survival time. Those patients are called left censored patients. This censoring is not very common and is not considered here.

The third kind is the interval censoring. It happens when the patient experiences the event between two examinations. This is often the case for non-lethal end-points like recurrence or appearance of distant metastasis. When the time between two examinations is sufficiently small, it may not require a specific processing. The time of the event is simply set at the end of the interval.

An important assumption on the censoring is made by most survival methods. It is assumed to be random and noninformative. As an example, let's consider a patient with a poor prognosis who needs some aggressive medications. This patient dies from a heart attack induced by their side effects and is censored. In this example, the censoring is in-

formative and non-random. The poor prognosis patient are more likely to be censored. Such situation should be avoided as much as possible.

These considerations are important. In retrospective data, however, they are difficult to control for the statistician. Like most survival methods, this thesis is focused on the standard survival analysis framework with non-informative right censored data.

2.2 Probabilistic view

The survival time t of a patient can be viewed as a realisation of a random variable T. The random variable T has an underlying probability density function $f(t)$ and cumulative distribution function $F(t)$.

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \tag{2.1}$$

$$F(t) = P(T \leq t) = \int_0^t f(u)\,\mathrm{d}u \tag{2.2}$$

$F(t)$ is the probability to have the event before time t. In survival analysis, it is often more convenient to work with $S(t)$ the survivor function. The survivor or survival function $S(t)$ is the probability of a patient to have a survival time greater than some value t. In cancer studies with distant metastasis as end-point, $S(t)$ is the probability of a patient to live free of metastasis until time t.

$$S(t) = P(T \geq t) = 1 - F(t) \tag{2.3}$$

2.2.1 Hazard function

Another very often used function is the hazard function $h(t)$. The hazard function is used to represent the risk or hazard of having the event at time t. The hazard function $h(t)$ is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.4}$$

It computes the probability density of a patient to experience the event between t and $t + \Delta t$, conditioned to his survival until time t. The hazard function is expressed in number of events per time unit and is not a probability. The hazard function can be expressed in terms of the probability density function and the survivor function:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.5}$$

From the previous equations (2.5), (2.3) and (2.2), we can derive that

$$h(t) = -\frac{d}{dt} \log S(t) \tag{2.6}$$

The probability density function $f(t)$ and the survival function $S(t)$ can also be expressed in terms of $h(t)$

$$S(t) = \exp\left(-\int_0^t h(u)\,\mathrm{d}u\right) \tag{2.7}$$

$$f(t) = h(t) \exp\left(-\int_0^t h(u)\,\mathrm{d}u\right) \tag{2.8}$$

With equation (2.8), the probability density function $f(t)$ can be computed from $h(t)$. In the special case where the hazard function is a constant $h(t) = 1/\lambda$, $f(t)$ follows an exponential distribution.

$$f(t) = \frac{1}{\lambda} \exp\left(-\frac{t}{\lambda}\right) \tag{2.9}$$

where $\lambda \in [0, \infty]$ is also known as the scale parameter of the exponential distribution. λ is also the mean of the distribution and thus the average survival time. The exponential distribution is a special case of the Weibull distribution which is often used as a simple model for survival data. The Weibull distribution is defined by two parameters: the scale λ and the shape k.

$$f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left(-(\frac{t}{\lambda})^k\right) \tag{2.10}$$

The corresponding hazard function is computed as:

$$h(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \tag{2.11}$$

The shape parameter k is used to define increasing or decreasing hazard function, resp. with $k > 1$ and $0 < k < 1$. When k equals 1 the Weibull distribution is an exponential distribution with a constant hazard function. This simple model is often used to generate survival data. The censoring is generated independently by another distribution, typically another Weibull.

2.3 Kaplan-Meier survival curve

A Kaplan-Meier survival curve is the representation of the Kaplan-Meier estimate of a survival function [77]. The Kaplan-Meier estimate is a nonparametric maximum likelihood estimator and the most widely used estimator of the survival function.

To understand how the estimation is made, let us consider a particular time point t_j where at least a patient experienced the event. The probability to survive through this time point t_j can be estimated from the number of patients at risk just before t_j and the number of patients who die at t_j, respectively n_j and d_j. A patient is at risk if he has not yet experienced the event and is not yet censored. The patient probability of surviving until $t_j + \Delta t$, knowing that he survived until t_j, is estimated by

$$\hat{P}(T > t_j + \Delta t | T \geq t_j) = \frac{(n_j - d_j)}{n_j} \tag{2.12}$$

Assuming all events to be independent, the survival function $S(t_j + \Delta t)$ or the probability to survive until $t_j + \Delta t$ can be expressed as

$$S(t_j + \Delta t) = P(T > t_j + \Delta t | T \geq t_j) P(T \geq t_j) \tag{2.13}$$

$$\hat{S}(t_j + \Delta t) = \frac{(n_j - d_j)}{n_j} S(t_j) \tag{2.14}$$

As a nonparametric estimator, no assumption is made on the underlying event distribution and hazard. The estimated probability to survive through a time interval without any observed event is then 1. $\hat{S}(t)$ is thus constant between any pair (t_j, t_{j+1}) of consecutive time points with occurring events.

$$\hat{S}(t_{j+1}) = \hat{S}(t_j + \Delta t) \tag{2.15}$$

From equations (2.14) and (2.15) , we can derive a general formulation for $\hat{S}(t)$ which includes all time points with an event prior to t.

$$\hat{S}(t) = \prod_{j|t_j \leq t} \left[\frac{(n_j - d_j)}{n_j} \right] \tag{2.16}$$

From $t = 0$ to the first event, $\hat{S}(t)$ equals 1. As censored patients are at risk until their censoring, they are included in all n_j prior to their censoring and thus contribute to the estimation. If there is no censoring, $(n_j - d_j)$ will be equal to n_{j+1} for two consecutive time points t_j and t_{j+1}. The estimate of the survival function $S(t)$ becomes simply the number of patients alive at t divided by the total number of patients.

A Kaplan-Meier survival curve is represented as a step function. Each step corresponds to the occurrence of at least an event. Traditionally, the censoring is marked with a small cross on the curve. Figure 2.1 is an example of Kaplan-Meier survival curve.

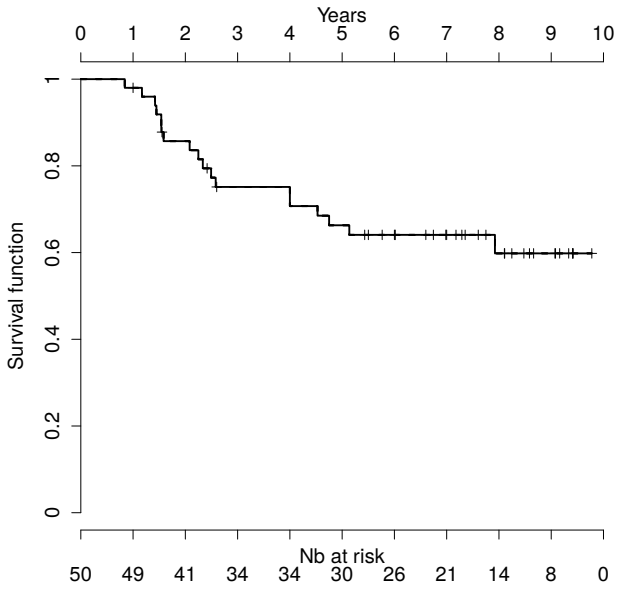


Figure 2.1: An example of Kaplan-Meier survival curve

2.4 Cox proportional hazards model

The Cox model is a regression model mainly used to estimate and predict the relative risk/hazard of patients. We consider each patient $i \in \{1, \ldots, n\}$ as being characterized by a 3-tuple $(t_i, \delta_i, \mathbf{x}_i)$, where t_i is the time of an event whenever δ_i equals 1 and the censoring time whenever δ_i equals 0. The vector $\mathbf{x}_i$ includes p covariates for the patient i. The covariates are variables that are possibly predictive of the survival, *e.g.* age, clinical factors, gene expression values, ...

The main idea behind the Cox model [30] is to express the hazard $h_i(t)$ of a patient i as the product of a baseline hazard $h_0(t)$ and a positive function of the covariates $\mathbf{x}_i$:

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x}_i) \tag{2.17}$$

The baseline hazard is shared among all patients. $r_i = \boldsymbol{\beta}^\top \mathbf{x}_i$ represents the patient relative risk, his risk score. The higher the risk score, the higher is the hazard $h_i(t)$ and the chances of experiencing an early event.

Modelling the hazard in this way makes the hazards $h_i(t)$ and $h_j(t)$ of any pair of samples (i, j) proportional. Their ratio $h_i(t)/h_j(t)$ is constant over time and do not depend on $h_0(t)$.

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\exp(\boldsymbol{\beta}^\top \mathbf{x}_j)} \tag{2.18}$$

The name of the model derives from this assumption of proportionality of the hazards. Its strength comes from the possibility to estimate the parameters $\boldsymbol{\beta}$ independently and without any further assumptions on $h_0(t)$. The Cox proportional hazards model is thus often referred to as a semi-parametric model. In practice, the relative risk of patients $r_i = \boldsymbol{\beta}^\top \mathbf{x}_i$ is more interesting than the actual hazard function $h_i(t)$. These risks can help clinicians to identify patients with poor prognosis who may need more attention or a particular treatment.

2.4.1 Partial likelihood

Along with the proportional hazard model, Cox proposed a way to fit the parameters $\boldsymbol{\beta}$ which is known as the partial likelihood maximization [30]. Except from the proportionality assumption, the method does not require the selection of a particular survival time distribution.

Let us now consider a particular time t_i where the patient i experienced the event. Conditioned on $R(t_i)$, the set of patients at risk at t_i, the probability that the event occurs for patient i as observed is given by

$$P(t_i|R(t_i)) = \frac{h_i(t_i)}{\sum_{k\in R(t_i)} h_k(t_i)} \tag{2.19}$$

$$P_{\boldsymbol{\beta}}(t_i|R(t_i)) = \frac{h_0(t_i)\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{k\in R(t_i)} h_0(t_i)\exp(\boldsymbol{\beta}^\top \mathbf{x}_k)} \tag{2.20}$$

$$P_{\boldsymbol{\beta}}(t_i|R(t_i)) = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{k\in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k)} \tag{2.21}$$

where $R(t_i) = \{k|t_k \geq t_i\}$ is the set of patients still at risk right before time t_i (still without event nor censoring). The likelihood $L_i(\boldsymbol{\beta})$ of the parameters for this event is given by $P_{\boldsymbol{\beta}}(t_i|R(t_i))$ and does not depend on the baseline hazard $h_0(t)$. The method relies only on the ranking of the survival times to compute $R(t_i)$ and is thus insensitive to any monotonic transformation of the time values. Assuming independence between events, the likelihood of $\boldsymbol{\beta}$ over all events is given by

$$L(\boldsymbol{\beta}) = \prod_{i|\delta_i=1} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{k\in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k)} \tag{2.22}$$

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{k\in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k)}\right]^{\delta_i} \tag{2.23}$$

It is worth noting that this likelihood is only exact when there is no ties in the event times. In such a case the conditional probabilities (2.19) of tied events should not be considered independently. However when there is not too many tied events, this simplification is a good approximation known as the Breslow approximation [21], which is used by default in many survival software.

This likelihood (2.23) is known as the partial likelihood. Partial in opposition to a full likelihood which takes into account the baseline hazard function not considered here. The maximization of this partial likelihood gives estimates of the parameters $\boldsymbol{\beta}$. Cox shows that these maximum partial likelihood estimates are consistent and asymptotically normal [31]. With an increasing number of patients, the estimates converge and are asymptotically unbiased. In many realistic situations, these estimators are also asymptotically fully efficient [40]. This means

that with a sufficient number of patients, the variances of the estimates will not be much larger than if we were using the full likelihood. Standard MLE inference methods can thus be used on the maximum partial likelihood estimators of the Cox model [31].

In chapter 3, more details are given on the methods used to maximize the partial likelihood. In particular when the number p of covariates is much larger than the number of patients n. In such a case, additional constraints are needed on the $\boldsymbol{\beta}$ to obtain good estimates.

Chapter 3

Modelisation and Prediction

An important aspect of cancer research is the development of better prognostic tools/models for clinicians. This chapter presents one of the key aspects of such tools: how they can be used to model and predict the possible outcomes of a disease.

The model predictions are based on a vector $\mathbf{x}_i$ with the p covariates of the patient $i \in \{1, \ldots, n\}$. In this thesis, those p variables are mostly gene expression data from microarrays. For simplicity, we assume that each variable x_j represents the level of expression of a different gene $j \in \{1, \ldots, p\}$. The number of genes is usually many times larger than the number of patients, $p \gg n$.

In cancer, the classical outcome of interest is the survival: the time to relapse, distant metastasis or death. As presented in section 2.2, the survival probabilities can be described through the hazard function. The objective of prognostic models is then to model and predict the hazard of patients, or at least their relative hazards/risks.

This relative risk can be continuous or discrete. The variable r_i represents a continuous risk, called the risk score, of a patient i. A patient with a higher risk score is more likely to have an early event than a patient with a lower risk score.

When the relative risk is discrete, it defines groups of patients with a similar outcome. These groups are called risk groups. In this thesis, we mainly use two risk groups defining the high and low risk patients. The predicted risk group g_i is either 1 or -1, respectively for the high and low risk groups.

This discretization in two (sometimes more) risk groups is a common practice in the medical literature. The risk groups are justified by the need of clinicians to make decisions *i.e.* whether or not a specific (and

sometimes aggressive) treatment should be given to a patient. Such decision is precisely based on the assignment of this patient to a particular risk group.

However, this discretization should only be applied to risk scores and not to continuous predictor variables. The variable dichotomization may indeed introduce problems such as loss of information, reduction in power, uncertainty in defining the cutpoint [110]. In this thesis, we only descretize risk scores at the very end of the model-building process. Moreover, the prognostic model and the discretization cut-off are always assessed on external validation data to avoid any optimistic bias.

In this thesis, we use two kinds of risk prediction methods using either survival or classification models. Their main difference is the nature of their supervisions, that is the external information available on the survival of patients or its true risk group. Unlike the covariates, this information is used to learn a model but is not available for the predictions on new patients.

The supervision used with survival models is the survival data, (t_i, δ_i) for each patient i. t_i is the time of an event whenever δ_i equals 1 and the censoring time whenever δ_i equals 0. With survival models, we assume that the true risk scores and risk groups of patients are unknown.

On the contrary, classification models are used precisely when the true risk groups are known. Here, the supervision is not some survival data but a variable y_i coding for the **true** risk group of a patient i. The definition of these group labels y_i is external and represents the decisions (or predictions) of clinicians. We differentiate g_i and y_i, predicted and true group labels, to make explicit when true labels are available at training time.

In this thesis, we mainly focus on (generalized) linear prognostic models for their relative simplicity and interpretability. This chapter starts with a general description of linear models, section 3.1, and motivates this specific choice of prognosis models. A particular focus is made on their use with high dimensional data such as microarrays. Section 3.2 presents two classification models namely the logistic regression and support vector machines. Section 3.3 revisits the Cox model presented in section 2.4 and shows its links with the logistic model. An extension of the Cox model using support vector machine is also presented.

3.1 Linear models

Linear models assume that the variable of interest Y can be described as a function f of a linear combination of the input variables (or covari-

ates) $\mathbf{x}$.

$$Y = f(\boldsymbol{\beta}^\top \mathbf{x}) \tag{3.1}$$

The linear combination of the variables, $r = \boldsymbol{\beta}^\top \mathbf{x}$, is named here the score of the model. The function f is used to link the score and the prediction. This function can be non-linear but it is fixed, not trained. The weights of the linear combination $\boldsymbol{\beta}$ are the model parameters. The challenge with linear prognostic models is to find the best parameters to have good prediction performances on new data.

The key advantage of linear models is their relative simplicity. Whatever method is used to fit the parameters $\boldsymbol{\beta}$, the model can be summarized as a weighted sum of the patients variables $\mathbf{x}$. The advantages of this simplicity are two folds.

First, the linear relation between the model parameters and the variables is easy to interpret. The absolute value of the parameter β_j, $\boldsymbol{\beta} = [\ \ldots\ \beta_j\ \ldots\ ]^\top$, can be interpreted as the importance of the variable x_j in the prognostic model. With the models presented in this chapter, the sign of a parameter further shows if a variable is positively or negatively correlated with the prognosis. Those parameters can thus be interpreted by clinicians and give them some insights on important risk factors. These risk factors can in turn be used to better understand the biological processes involved in the recurrence or metastatisation. The simplicity and interpretability make the linear models understandable by clinicians. It is particularly important, if we want them to trust and use prognostic models.

The second advantage comes from the data used in this thesis: microarrays. The microarrays are used here to measure the expression levels of thousands of genes which are used as potential prognostic markers. Due to their price, these microarrays data are not available in quantity. The number of genes is usually many times larger than the number of patients, $p \gg n$. In such a setting, finding a good set of model parameters $\boldsymbol{\beta}$ becomes very difficult. This problem is known as the curse of dimensionality. Allowing more complex models with non-linear relationships in the variables will increase even more the difficulties. These complex model are then very often outperformed by simpler linear ones.

3.1.1 Fitting the parameters

Fitting the parameters of the model can be viewed as an optimization problem. The solution of this optimization problem is a set of parameters $\hat{\boldsymbol{\beta}}$ minimizing a particular function of the parameters.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\beta}) + \lambda \mathcal{R}(\boldsymbol{\beta}) \tag{3.2}$$

where $\mathcal{L}$ is known as the loss of the model and $\mathcal{R}$ is a regularization term. λ is a meta-parameter controlling the relative importance between loss and regularization. This pattern *loss + regularization* is very common with linear model for high dimensional data such as microarrays. In particular, the pattern fits with all models presented here.

Loss function

The loss function $\mathcal{L}(\boldsymbol{\beta})$ defines the fit of the model to the observed data. The better the fit, the lower the loss. The loss measures in particular how far are the observations (survival times, group labels, etc) from the model predictions. This "distance" between observations and predictions changes with the assumptions of each model, leading to many different losses. In this chapter, the models are thus presented and compared through their loss functions.

Regularization

When the number of patients n is smaller than the number of variables p, there are many solutions $\hat{\boldsymbol{\beta}}$ minimizing the loss function. In such a setting with $p \gg n$, the problem is then how to choose a good one. Following the Occam's razor principle, among these apparently equal solutions the simplest one is the better.

If the loss function $\mathcal{L}(\boldsymbol{\beta})$ defines how a model fits the data, the regularization $\mathcal{R}(\boldsymbol{\beta})$ computes how complex the model is. Adding a regularization $\mathcal{R}(\boldsymbol{\beta})$ in the optimization problem (3.2) is thus a way to penalize complex models and favour the simplest ones.

The two most common regularization are known as the L_1 and L_2 regularizations. The L_2 regularization, or ridge penalty [65], is the square of the Euclidean (L_2) norm of $\boldsymbol{\beta}$. The L_1 regularization, or lasso [132], is the Manhattan (L_1) norm.

$$L_2 = \frac{1}{2} \sum_{j=1}^{p} \beta_j^2 = \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 \tag{3.3}$$

$$L_1 = \sum_{j=1}^{p} |\beta_j| = \|\boldsymbol{\beta}\|_1 \tag{3.4}$$

In a probabilistic setting, the regularization can be viewed as a prior (an assumption) on the distribution of the parameters [53]. Minimizing the pattern *loss + regularization* will then find the parameters $\boldsymbol{\beta}$ that maximize the likelihood of both the observations and the parameters. The L_1 and L_2 regularizations respectively correspond to a Laplace and a Gaussian prior. The meta-parameter λ can be interpreted as a way to define the variance of the parameters distribution. The higher is λ, the lower is the variance of the prior distribution.

The L_1 norm is a sparsity inducing norm. This regularization is particularly efficient when there are only few relevant variables [132]. However, it can sometimes lead to overly sparse models. In particular, the L_2 regularization outperforms the lasso when there is a high correlation between the variables.

Zou and Hastie [147] introduce the elastic-net penalty to combine the strength of these two regularizations. The elastic-net penalty is a combination of the L_1 and L_2 penalty.

$$L_e = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 \tag{3.5}$$

$$= \left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2\right) \tag{3.6}$$

where $\alpha \in [0-1]$ is a meta-parameter of the regularization. The elastic-net regularization is reduced to the lasso (respectively to ridge penalty) whenever $\alpha = 1$ (respectively $\alpha = 0$).

The lasso and elastic-net regularizations can be seen as some form of soft-thresholding of the parameters Zou and Hastie [147], Tibshirani [132]. Using these regularizations will then shrink many parameters to 0. These penalties are thus a way to select relevant variables and to improve the interpretability of prognostic models. This property of some regularizations is discussed in the chapter 5 with other feature/variable selection techniques.

3.2 Classification models

Classification models are used to model and predict risk groups. In this section, we assume that the true risk groups are known at training time and can be used when estimating the model. The definition of these group labels y_i is external and represents the decision process of clinicians. This decision can be based on the observed survival times, risk factors or cancer subtypes.

As an example, the low risk patients $y_i = -1$ can be defined as the patients without events or that live longer than a particular time threshold. In both cases, we have problems with early censoring. Those early censored patients cannot be safely assumed as neither at high nor at low risk. Another problem is how to choose a good time threshold. In practice, this approach is not very common to learn a classification model such as a logistic regression or a SVM. A survival model using the full survival data is preferred in this thesis when risk groups are not directly available.

The classification models are more useful when risk groups are truly observed, for example with cancer subtypes or risk factors. We then have a real classification problem where the groups have differences in their survival. An example of such classification problem is the prediction of the histologic grade of the tumor which is highly correlated with the survival. A good classification model can improve the histological grading (or replace it) and help survival prediction.

We present in sections 3.2.1 and 3.2.2 logistic regression and support vector machines. They are two very common classification models and have a strong connection with the two survival models presented in section 3.3. The classification schemes for multi-class problems are not covered here, but extensions exist along the same lines [12].

3.2.1 Logistic regression

The logistic model is a binary classification model. It assumes the risk group of a patient i to be computed as the sign of an unknown risk score, r_i^*. This unknown risk score r_i^* can be written as a linear combination of the patients covariates $\mathbf{x}_i$ in addition to a random error term ϵ.

$$y_i = \text{sign}(r_i^*) \tag{3.7}$$
$$= \text{sign}(\boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon) \tag{3.8}$$

where ϵ is distributed according to a standard logistic distribution. The high risk patients are the patients with a positive risk score $r_i^* = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon$. The probability of a patient i to be in the high risk group (the positive class) is given by

$$P(Y_i = 1|\mathbf{x}_i) = P(r_i^* > 0) \tag{3.9}$$
$$= P(\boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon > 0) \tag{3.10}$$
$$= P(-\epsilon < \boldsymbol{\beta}^\top \mathbf{x}_i) \tag{3.11}$$

The logistic distribution is symmetric. The $-\epsilon$ can be replaced by ϵ. This probability can be computed from the cumulative distribution function of a logistic random variable.

$$P(Y_i = 1|\mathbf{x}_i) = P(\epsilon < \boldsymbol{\beta}^\top \mathbf{x}_i) \tag{3.12}$$

$$= \frac{1}{1 + \exp(-\boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{3.13}$$

The probability of patient i to be in the negative class can be computed from the previous equation.

$$P(Y_i = -1|\mathbf{x}_i) = 1 - P(Y_i = 1|\mathbf{x}_i) \tag{3.14}$$

$$= 1 - \frac{1}{1 + \exp(-\boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{3.15}$$

$$= \frac{\exp(-\boldsymbol{\beta}^\top \mathbf{x}_i)}{1 + \exp(-\boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{3.16}$$

$$= \frac{1}{1 + \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{3.17}$$

The probability of a patient i to be in his observed class/group can be computed as follow:

$$P(Y_i = y_i|\mathbf{x}_i) = \frac{1}{1 + \exp(-y_i \boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{3.18}$$

The likelihood of the parameters $\boldsymbol{\beta}$ is computed as the product of the probability of the observed labels in the training set.

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{1 + \exp(-y_i \boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{3.19}$$

The optimal model parameters are learned by maximizing this likelihood. Equivalently, we can minimize minus its logarithm, the loss function of the logistic model.

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{\beta}^\top \mathbf{x}_i)) \tag{3.20}$$

In the logistic model, the score $r_i = \boldsymbol{\beta}^\top \mathbf{x}_i$ can be used to compute and to predict risk group probabilities. Assigning new patients to their most likely risk group gives us the following decision function:

$$g_i = \text{sign}(\boldsymbol{\beta}^\top \mathbf{x}_i) \tag{3.21}$$

In high-dimension, the logistic model needs some regularization: L_1, L_2 or elastic-net. Efficient algorithms exist to solve the optimization problem of the regularized logistic model. They often relies on an iteratively reweighted least squares methods such as the method proposed by Friedman *et al.* [52].

3.2.2 Support vector machines

Support vector machines, or SVM, are binary classification models. Similarly to the logistic model, the decision function of an SVM is defined as:

$$g_i = \text{sign}(\boldsymbol{\beta}^\top \mathbf{x}_i) \tag{3.22}$$

but the estimation procedure for the $\boldsymbol{\beta}$ differs. The parameters $\boldsymbol{\beta}$ can be viewed as the definition of a hyperplane. A hyperplane is the generalization of a 3D plane for an arbitrary number of dimensions. Here, we have a plane in a p-dimensional space defined by the p covariate of the patients. The hyperplane is defined as the set of points for which $\boldsymbol{\beta}^\top \mathbf{x} = 0$. The decision function of both an SVM and a logistic regression can be interpreted as looking on which side of the hyperplane are the samples/patients.

The definition of an SVM comes from the question: what is a good hyperplane to classify new unseen samples? One reasonable solution is to choose the hyperplane achieving the largest separation between the two classes/groups. In other words, the hyperplane that maximize the distance with the closest points of both sides. This distance is called the margin. These closest points are defined as the samples $\mathbf{x_i}$ such that

$$|\boldsymbol{\beta}^\top \mathbf{x_i}| = 1 \tag{3.23}$$

The size of the margin can be computed as $1/\|\boldsymbol{\beta}\|_2$. Finding the best hyperplane becomes a minimization problem:

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 \tag{3.24}$$

under the constraints that every sample is outside the margin and on the correct side of the hyperplane. These constraints are written:

$$y_i \boldsymbol{\beta}^\top \mathbf{x_i} \geq 1 \quad \forall i \in 1, \ldots, n \tag{3.25}$$

With nonlinearly separable data, the margin constraints are too strong. A soft-margin version of SVMs was introduced by Cortes and Vapnik [29] to solve this issue. The problem becomes the minimization of both $\|\boldsymbol{\beta}\|_2^2$ and the violation of the margin constraints.

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C \sum_{i=1}^{n} \left[1 - y_i \boldsymbol{\beta}^\top \mathbf{x_i}\right]_+ \tag{3.26}$$

where the function $[x]_+$ is defined as

$$[x]_+ = \begin{cases} x & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases} \tag{3.27}$$

A sample contribute in the SVM minimization only if he violates the constraint, that is when $y_i \boldsymbol{\beta}^\top \mathbf{x_i} \leq 1$. Those contributing samples are called the support vectors. Usually, the number of support vectors is small. The SVM is thus robust to small changes of the data set that keep the support vectors untouched.

Defining C as $1/(2\lambda)$, this formulation of the soft-margin SVM perfectly fit in the *loss + regularization* pattern. The SVM loss is known as the hinge loss and is close to the logistic loss, as shown in figure 3.1. The SVM is thus very similar to a L_2 regularized logistic regression but does not offer a probabilistic interpretation of its output. Rosset *et al.* [107] even prove that under some conditions the logistic regression and SVM converge to the same solution.

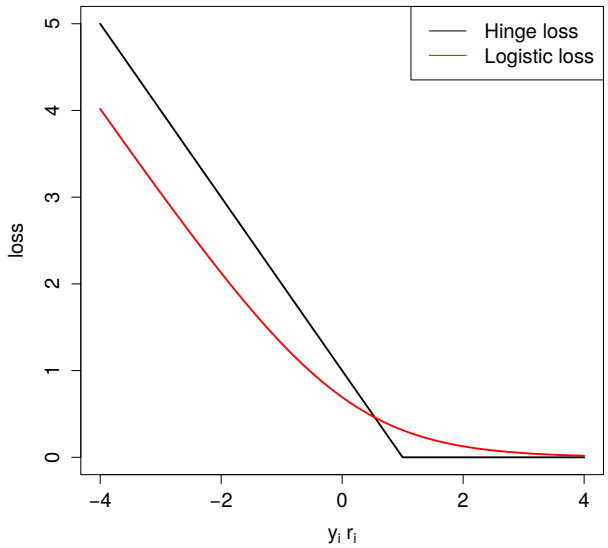


Figure 3.1: Comparison between hinge and logistic losses

3.3 Survival models

The objective of linear survival models presented in this section is to model and predict the survival of patients through their risk scores.

$$r_i = \boldsymbol{\beta}^\top \mathbf{x}_i \tag{3.28}$$

Unlike the classification models, survival models do not have assumption or information on the risk groups. The model parameters $\boldsymbol{\beta}$ are learned using only the survival data (t_i, δ_i). The risk score of the survival model cannot be directly used to compute the risk group. An additional cutoff θ should be provided to define the decision function:

$$g_i = \text{sign}(\boldsymbol{\beta}^\top \mathbf{x}_i - \theta) \tag{3.29}$$

$$g_i = \begin{cases} 1 & \text{if} \quad r_i > \theta \\ -1 & \text{if} \quad r_i \leq \theta \end{cases} \tag{3.30}$$

There are many possible choices for the parameter θ. It can be defined to fit the expected proportion of risk groups or to maximize some criteria on the quality of those groups. Such criteria are presented

in the chapter 4. The effect of the specific choice of parameter θ on the risk groups is discussed in chapter 8.

The most common survival model is the Cox proportional hazards model presented in section 2.4. Section 3.3.1 presents how it can be used with high dimensional data and its links to the logistic model. Section presents an extension of the Cox model using SVMs, which have the advantage of being sparse in the number of observations used.

In this thesis, we focus mainly on linear models but other kinds of survival models exist such as the random survival forests [71], the partial logistic artificial neural networks [11], *etc.*

3.3.1 Cox regression

As presented in details in section 2.4, the Cox model [30] assumes that the hazard $h_i(t)$ of a patient i is the product of a baseline hazard $h_0(t)$ and a positive function of the covariates $\mathbf{x}_i$:

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x}_i) \tag{3.31}$$

The estimation of the parameters $\boldsymbol{\beta}$ can be done by maximizing a partial likelihood.

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{k \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k)} \right]^{\delta_i} \tag{3.32}$$

where $R(t_i) = \{k | t_k \geq t_i\}$ is the set of patients still at risk right before time t_i (still without event nor censoring). To fit in the *loss + regularization* pattern, the likelihood can be turned into a loss function defined as minus its logarithm.

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[-\boldsymbol{\beta}^\top \mathbf{x}_i + \log \sum_{k \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k) \right] \tag{3.33}$$

An additional regularization term can be added to the minimization process to avoid overfitting. Like the logistic regression, the cox model is a generalized linear model. The same efficient algorithms can be used to solve both the Cox and Logistic regularized minimization problems [52, 117].

Cox model and logistic regression

Section 2.4 presents the classical interpretation of the Cox model leading to the partial likelihood. We present here an alternative interpretation of the Cox model using the conditional logistic model [27, 28]. The links between conditional logistic model and Cox model is not always well known but is described in [24], which is the source of inspiration for this section. This interpretation is used later in section 3.3.2 and chapter 6.

As a consequence of the proportional hazard assumption, no information can be used from the time intervals between consecutive events [30]. The parameters $\boldsymbol{\beta}$ of the cox model are thus learned by looking at all time points with events. Each time point is characterized by a set of observed events among patients at risk, which can thus be viewed as a binary classification problem. The positive class contains the events. The negative class contains the patients censored or surviving further. The conditional logistic model allows us to compute the probabilities of these particular classes:

$$P(Y_i = 1|\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{3.34}$$

$$P(Y_i = -1|\mathbf{x}_i) = \frac{1}{1 + \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{3.35}$$

Those probabilities can be used at a particular time point t_i where patient i experienced the event. The conditional probability of this event knowing that exactly one event occurs and the set of patients at risk is given by:

$$P(t_i|R(t_i), \mathbf{x}_i) = \frac{P(Y_i = 1|\mathbf{x}_i) \prod_{j \in R(t_i)\setminus\{i\}} P(Y_j = -1|\mathbf{x}_j)}{\sum_{k \in R(t_i)} P(Y_k = 1|\mathbf{x}_k) \prod_{j \in R(t_i)\setminus\{k\}} P(Y_j = -1|\mathbf{x}_j)} \tag{3.36}$$

The numerator in equation (3.36) is decomposed as the probability of patient i having the event $P(Y_i = 1|\mathbf{x}_i)$ and all the others patients, among those still at risk at time t_i (*i.e.* in $R(t_i)$), not experiencing this event ($P(Y_j = -1|\mathbf{x}_j)$). As a conditional probability, the denominator is the normalization factor representing the probability of having exactly one event, not specifically for patient i, among the set $R(t_i)$ of patients at risk at time t_i. This expression is thus similar to the numerator but with a summation and a running index k over the set of patients still at risk.

Such a conditional logistic regression is also used in matched case-control studies [27] where specific patients of interest are distinguished from control patients in a classification context. Here, in a survival context, we consider in equation (3.36) a single patient of interest (the one experiencing the event at time t_i) and all other patients as control cases. In other words, equation 3.36 can be considered analogous to a 1:M matched case-control study with one case and M controls (see equation (B.6) of appendix B.2 from [27]).

Combining equation (3.36) with the class probabilities (3.34) and (3.35) leads to:

$$P(t_i|R(t_i),\mathbf{x}_i) = \frac{\frac{\exp(\boldsymbol{\beta}^\top\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}^\top\mathbf{x}_i)}\prod_{j\in R(t_i)\setminus\{i\}}\frac{1}{1+\exp(\boldsymbol{\beta}^\top\mathbf{x}_j)}}{\sum_{k\in R(t_i)}\frac{\exp(\boldsymbol{\beta}^\top\mathbf{x}_k)}{1+\exp(\boldsymbol{\beta}^\top\mathbf{x}_k)}\prod_{j\in R(t_i)\setminus\{k\}}\frac{1}{1+\exp(\boldsymbol{\beta}^\top\mathbf{x}_j)}} \tag{3.37}$$

$$= \frac{\exp\left(\boldsymbol{\beta}^\top\mathbf{x}_i\right)}{\sum_{k\in R(t_i)}\exp\left(\boldsymbol{\beta}^\top\mathbf{x}_k\right)} \tag{3.38}$$

Equation (3.38) is exactly the partial likelihood of the event as defined previously. It is worth noticing that the solution is invariant by translation, as opposed to a logistic regression solution. If we replace $\boldsymbol{\beta}^\top\mathbf{x}_i$ by $\boldsymbol{\beta}^\top\mathbf{x}_i-\theta$, the likelihood is unchanged as any additional θ will be cancelled out in equation (3.38). Without a θ, the risk scores $r_i = \boldsymbol{\beta}^\top\mathbf{x}_i$ may not be well centered and cannot be interpreted in class probabilities through equations (3.34) and (3.35).

The Cox model can thus be seen as a set of classification models, one for each time with at least an event, sharing the same parameters $\boldsymbol{\beta}$ but with different decision functions.

3.3.2 Survival-SVM

Survival-SVMs form a natural extension of Cox models proposed in Van Belle *et al.* [136], Evers and Messow [44]. Survival-SVMs are based on the interpretation of a Cox model as a set of classification models. The conditional logistic models are replaced here by SVMs sharing the same parameters $\boldsymbol{\beta}$. Using SVMs has the advantage of producing sparse models in the observation space.

Survival-SVM minimizes $\|\boldsymbol{\beta}\|_2^2$ under a set of constraints defining the margin, similarly to the standard SVM. For each classification problem, the margin is defined as the space between the patients experiencing the event and the patients who survive further. Those constraints can be written in terms of differences between pairs of patients.

$$\boldsymbol{\beta}^\top \mathbf{x}_i - \boldsymbol{\beta}^\top \mathbf{x}_k > 1 \tag{3.39}$$

$$\boldsymbol{\beta}^\top (\mathbf{x}_i - \mathbf{x}_k) > 1 \tag{3.40}$$

for all pairs i, k such that i experienced the event at time t_i and k survived at t_i. A soft-margin version of a survival-SVM is obtained by relaxing those constraints. The minimization problem becomes:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i|\delta_i=1} \sum_{k \in R(t_i^+)} \left[1 - \boldsymbol{\beta}^\top (\mathbf{x}_i - \mathbf{x}_k)\right]_+ + \frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2 \tag{3.41}$$

where $R(t_i^+) = \{k | t_k > t_i\}$ is the set of patients surviving longer than patient i. The loss of a survival-SVM is a convex approximation to the concordance index [62] presented in chapter 4. The concordance index is a popular metric to assess the performance of a survival model. The concordance index computes the probability of patients to experience the event before patients with a lower risk.

Unlike with a Cox model, a survival-SVM risk score cannot be interpreted in terms of hazard and is given by:

$$r_i = \boldsymbol{\beta}^\top \mathbf{x}_i \tag{3.42}$$

By using the same parameters $\boldsymbol{\beta}$ for all time points, a survival-SVM makes implicitly the proportional hazards assumption. Without this assumption, the relative risks of patients could change over time which is not the case here. All risk scores r_i are constant over time.

Evers and Messow [44] make an interesting comparison between the unregularized loss functions of a survival-SVM and a Cox model. They consider a particular situation with three patients: patient 1 experienced the event, patients 2 and 3 are censored later. With the risk scores $r_i = \boldsymbol{\beta}^\top \mathbf{x}_i, \quad \forall i \in \{1, 2, 3\}$, the loss of a survival-SVM becomes:

$$[1 - (r_1 - r_2)]_+ + [1 - (r_1 - r_3)]_+ \tag{3.43}$$

Similarly to a survival-SVM, the loss of a Cox model can be expressed in terms of differences between the risk score of patients. From equation (3.33), the loss of a Cox model becomes:

$$\log(1 + \exp(-(r_1 - r_2)) + \exp(-(r_1 - r_3))) \tag{3.44}$$

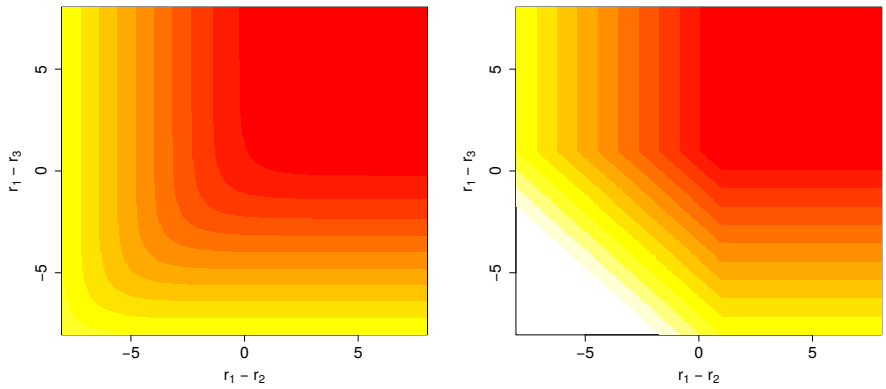


Figure 3.2: Losses of a Cox and a Survival-SVM in an example with three patients.

Both losses tend to 0 when $r_1 - r_2$ and $r_1 - r_3$ are going to $+\infty$. When $r_1 - r_2$ is sufficiently large, the losses are the logistic and SVM losses with respect to $r_1 - r_3$. Figure 3.2 illustrate the two losses while changing the differences between the three risk scores. The similarities between a survival-SVM and a regularized Cox model explain their similar results as reported in [136, 44, 138, 137].

Chapter 4

Performance Assessment

This chapter defines different metrics used to assess the performances of the prognostic models presented in chapter 3. The performance metrics are divided in two categories: the metrics to assess the classification performances (section 4.1) and those to assess the survival prediction (section 4.2). These metrics describe how well the model predictions fit the observed true risk groups or times of events, respectively for the classification and survival prediction.

The prediction tasks in this thesis are mostly binary: high versus low risk groups. The metrics are thus presented in this context. Most of them are however easily extended to multi-group settings.

4.1 Classification performances

The classification models assume the risk groups are known at training time. The risk groups represent the decision of clinicians, defining patients either at high or low risk. In this section, we present metrics to assess how good are the model predictions with respect to clinician decisions.

The classification results can be summarized in a confusion matrix, such as table 4.1 for a binary classification problem. A confusion matrix contains a row and a column for each possible risk group. Each column, resp. row, represents a true, resp. predicted, class/group label. A table cell contains the number of patients with a particular combination of true and predicted risk groups.

Risk groups		**True**	
		High	Low
Predicted	High	TP	FP
	Low	FN	TN

Table 4.1: Confusion matrix

4.1.1 Accuracy

The accuracy is the proportion of patients predicted with the correct label. Classification accuracy is one of the most commonly used metrics to summarize the confusion matrix presented in table 4.1:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

In a multi-classes setting, the accuracy can be generalized as the sum of the diagonal of the confusion matrix divided by the total number of samples. A major problem of this metric appears with unbalanced data set. An unbalanced data set is a data set where the samples are not distributed equally in all classes. One or several classes are overrepresented. In this context, the importance of a class in the accuracy is proportional to its size. If 90% of the samples are in one class, the prediction performance in this class account for 90% of the global accuracy. A naive classifier giving to every samples the label of this majority class will have 90% of accuracy. This problem is addressed by the BCR defined in section 4.1.3.

4.1.2 Sensitivity and specificity

In some cases, the cost of a classification error is not equal for each class. Failing to identify a high risk patient can have very bad consequences, as he may need some particular treatments. The opposite, wrong labelling of a low risk patient may be discovered with further tests. In the medical domain, the model accuracies in each class are thus often reported. For a binary classification problem, they are called the sensitivity and specificity.

Sensitivity, or true positive rate, is the proportion of patients at high risk correctly identified as such. It answers the question: are we able to find the high risk patients?

$$\text{SE} = \frac{TP}{TP + FN} \tag{4.2}$$

Specificity, or true negative rate, is the proportion of samples in the negative class (low risk) identified as such.

$$\text{SP} = \frac{TN}{TN + FP} \tag{4.3}$$

A classification model generally offers a particular trade-off between those two aspects. The sensitivity can generally be increased at the cost of a poor specificity. They are thus not informative alone and should be reported together.

4.1.3 Balanced classification rate

The balanced classification rate (BCR) is the arithmetic average of the accuracy per class label. With two classes, it is the mean between sensitivity and specificity.

$$BCR = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{4.4}$$

When the classes are of the same size, the BCR is equivalent to the Accuracy (4.1.1). The BCR overcomes the problem of unbalanced data accuracy. A naive classifier which gives the same label to every samples always has a BCR of 50% regardless of the class sizes. Unlike the accuracy, the expected BCR of a random classifier is always 50% for any true and predicted class priors.

4.1.4 Area under the ROC curves

The area under the ROC curves differs from the previous classification metrics. It assess the performance of the score (the continuous output) of a classifier without considering its decision threshold.

The ROC curve depicts the trade-offs between sensitivity and specificity while changing the decision threshold. The ROC curve is plotted in the 2-dimensional ROC space defined by the sensitivity and 1-specificity. Each possible decision threshold defines a particular classification with a specificity and sensitivity and is a point of the ROC curve. An example of ROC curve is presented in figure 4.1.

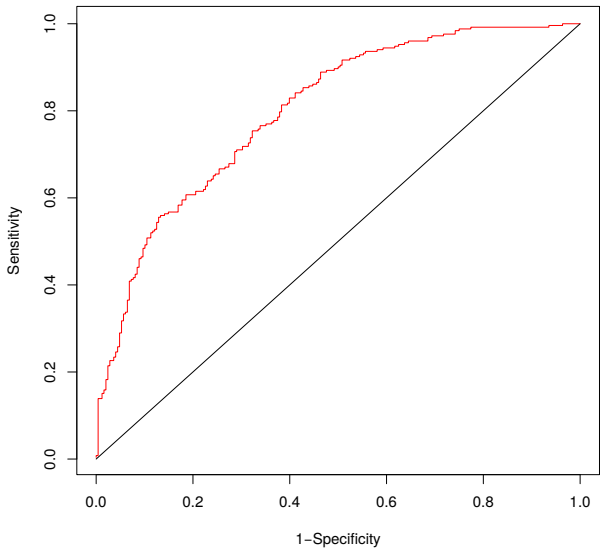


Figure 4.1: An example of ROC curve

The ROC space is a unit square as the sensitivity and the specificity are defined over $[0-1]$. The top-left corner equals to a specificity and sensitivity of 1 and thus a BCR of 1. At the opposite, the bottom-right corner corresponds to a classifier with a BCR of 0. The diagonal between the two last corners is an isoline corresponding to a BCR of 0.5. All parallels to this diagonal are also isolines of BCR. A good ROC curves should thus be as far as possible from the diagonal.

The AUC, Area Under the ROC Curve, is used to summarize the ROC curve in a single number. The AUC can be viewed as a probability. For a random pair of samples from each class, the AUC is the probability of the continuous output (risk score) of the classifier to have a higher value for the sample of the positive class (at high risk). An AUC of 1 means that the classifier achieves a perfect ranking of the patients. 0.5 is the AUC expected from a random model.

4.2 Survival prediction performances

The metrics, presented in this section, compute how well the survival model predictions fit the observed times of events. As presented in chapter 3, survival models predictions are of two kinds: risk score r_i (continuous) or risk group g_i (discrete). With the survival model used

in this thesis, the risk score is defined as a linear combination of the patient variables:

$$r_i = \boldsymbol{\beta}^\top \mathbf{x}_i \tag{4.5}$$

The predicted risk groups are defined from the risk scores as:

$$g_i = \begin{cases} 1 & \text{if} \quad r_i > \theta \\ -1 & \text{if} \quad r_i \leq \theta \end{cases} \tag{4.6}$$

The risk group g_i of a patient i is 1 for the high risk group and -1 for the low risk group. The discretization in two (sometimes more) risk groups is a common practice in the medical literature. The groups are used to make decisions on the treatments given to a patient. A proper assessment of the risk groups is then particularly important. The performance metrics for survival are presented here in this context of binary risk group prediction. However, most of them can be easily extend to assess risk scores or multiple risk groups.

4.2.1 Hazard ratio

The group hazard ratio (HR) evaluates the difference between survival curves computed by a Cox proportional hazards model (see section 2.4). It represents the increase in the risk of event between the low and high risk groups. When used to evaluate risk groups, the hazard ratio is computed with a Cox model using the binary group variable g_i as single covariate. Usually, the two risk groups are defined with 0 and 1. To have the same hazard ratio definition using -1 and 1, we divide g_i by two in the Cox model. The hazard function $h_i(t)$ for a patient i is then written as:

$$h_i(t) = h_0(t) \exp\left(\frac{\beta g_i}{2}\right) \tag{4.7}$$

Since g_i equals -1 or 1 for the patients in respectively the low risk or high risk group, the hazard of each risk group is given by:

$$h_{\text{High}}(t) = h_0(t) \exp\left(\beta/2\right) \tag{4.8}$$

$$h_{\text{Low}}(t) = h_0(t) \exp\left(-\beta/2\right) \tag{4.9}$$

The hazard ratio HR is the ratio between the hazard of the two risk groups.

$$HR = \frac{h_{\text{High}}(t)}{h_{\text{Low}}(t)} \tag{4.10}$$

$$= \frac{h_0(t)\exp(\beta/2)}{h_0(t)\exp(-\beta/2)} = \exp(\beta) \tag{4.11}$$

With the proportional hazard assumptions [30], $h_0(t)$ vanishes in the estimation of the hazard ratio. The HR, $\exp(\beta)$, is then constant over time and represents the increased risk of being at high risk with respect to the low risk patients. The higher the hazard ratio, the higher the difference between the two hazard functions and thus between the survival curves.

The hazard ratio is computed as the solution of fitting a Cox model with only one covariate g_i and one parameter β. The log-hazard ratio β maximizes the partial likelihood and has most of the good properties of maximum likelihood estimator (see section 2.4.1). With only one variable to compute the hazard ratio, the parameter of the Cox model is written β and not $\boldsymbol{\beta}$. In this thesis, we will use β when a hazard ratio is computed and $\boldsymbol{\beta}$ for multivariate prognostic models.

The hazard ratio can also be computed for more groups or a continuous risk score. In the continuous case, g_i is replaced by the risk score r_i in the equations.

4.2.2 Concordance index

The concordance index (C-index) measures to which extent the risk groups are concordant with the time to event, that is whether the patients in the high risk group actually experience the event before the patients in the low risk group [62].

The C-index specifically relies on the notion of comparable pairs. A pair of patients $\{i, k\}$ is comparable if patient i experiences the event while the patient k is still at risk (not censored and not having experienced the event) at time t_i. Such comparable pair is concordant if the patient i, experiencing the event earlier, belongs to the high risk group and patient k in the low risk group, $g_i > g_k$. The C-index lies between 0 and 1 as it estimates the probability for a comparable pair of patients to be concordant. This estimate is the number of concordant comparable pairs divided by the number of comparable pairs:

$$\text{C-Index} = \frac{\sum_{i,k\in\Omega} \text{Conc}(g_i, g_k)}{|\Omega|} \tag{4.12}$$

with

$$\text{Conc}(g_i, g_k) = \begin{cases} 1 & \text{if} \quad g_i > g_k \\ 0.5 & \text{if} \quad g_i = g_k \end{cases} \tag{4.13}$$

$$\Omega = \{(i,k) | \delta_i = 1, t_i < t_k\} \tag{4.14}$$

Ω is the set with all comparable pairs of patients. For the pairs of patients in the same groups, $\frac{1}{2}$ rather than 1 is added to the count of concordant pairs. The C-index can be directly used in the continuous case by replacing the groups g_i by r_i the risk score. Furthermore, it is not limited to survival data as the notion of comparable pairs can be extended to other response variables.

Links to classification metrics

In a classification setting, the set of comparable pairs is defined as the set of patient pairs (i, k) such that i and k are respectively in the high and low true risk groups.

$$\Omega = \{(i,k) | y_i = 1, y_k = -1\} \tag{4.15}$$

With this definition, computing the C-index with the risk scores of a classifier gives the AUC (section 4.1.4) as they estimate the same probability. Furthermore in this context, the C-index of the predicted risk groups g_i is the BCR presented in section 4.1.3. This last result can be easily found from equation (4.12) and using the confusion matrix in table 4.1.

4.2.3 Logrank test

The logrank is the statistics of a test to assess whether there is a significant survival difference between risk groups [88]. For each time point t_i where at least one patient has the event, the numbers of patients in each group and having the event can be summarized in a table:

Risk Groups	Number of events at t_i	Number surviving beyond t_i	Number at risk just before t_i
I (High)	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
II (Low)	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Table 4.2: Number of events for the two groups

Assuming the numbers of patients in each group (n_{1i} and n_{2i}) and having the event d_i as fixed, table 4.2 can be solely determined by d_{1i}. Under the null hypothesis that there is no survival difference across groups, d_{1i} follows an hypergeometric distribution. The expected value of d_{1i} under this null hypothesis, e_{1i}, is the mean of this hypergeometric distribution which can be computed as:

$$e_{1i} = \frac{d_i n_{1i}}{n_i} \tag{4.16}$$

The logrank statistic measures differences during the follow-up in number of events between what is observed d_{1i} in a group and what is expected under the null hypothesis e_{1i}. It is computed over all time t_i with an observed event.

$$U = \sum_i (d_{1i} - e_{1i}) \tag{4.17}$$

A high logrank implies that there is a higher than expected rate of events in the high risk group and more evidence against the null hypothesis. The logrank statistics is a sum of hypergeometric random variables (one variable for each time step t_i with an event) which approximately follows a normal distribution, from which a p-value can be easily computed.

4.2.4 SEP

The SEP metric compares the risk of each risk group with the risk of the entire population. SEP is a weighted geometric mean of the absolute relative risks between the risk groups and the global population [112].

$$SEP = \exp\left(\frac{n_{\text{High}}|\beta_{\text{High}}| + n_{\text{Low}}|\beta_{\text{Low}}|}{n}\right) \tag{4.18}$$

n_{High} and n_{Low} are the number of patients in respectively the high and low risk groups. β_{High} and β_{Low} are the log hazard ratios between each group and the whole population. SEP is not designed to handle continuous risk scores but can be easily extended to many risk groups.

$$SEP = \exp\left(\sum_k \frac{n_k}{n}|\beta_k|\right) \tag{4.19}$$

where n_k is the number of patients in the k^{th} risk group, and β_k is the log hazard ratio between the k^{th} risk group and the whole population. The SEP metric is designed to be interpretable in a similar way as the hazard ratio. It computes the average multiplicative factor between the hazard of a risk group and the global hazard of the population. To be clinically useful, a risk group needs to be well separated but also to incorporate a substantial part of the patients [112]. This particular point is assessed while introducing an average weighted by the risk group sizes. If one particular risk group contains nearly all patients, its hazard should be close to the global hazard and the corresponding β_k close to 0. The SEP metrics will then be close to 1 which is the minimal value of the metric.

Royston and Sauerbrei [108] proposed an extension of the SEP metric the D-index. The D-index behaves very similarly to a hazard ratio with a small correction for unbalanced risk groups. Unlike SEP, it can be used for the assessment of risk scores and risk groups.

4.2.5 Sensitivity and specificity

In some cases, it is useful to investigate the survival at a particular time threshold. For instance, in the context of breast cancer prognosis, 5 years after treatment is commonly accepted as a critical value from a clinical viewpoint. Defining such a time threshold transforms the survival prediction in a classification problem. The goal is to predict the patients who will experienced the event before this threshold or live longer. The risk group predictions can be assessed with the classification metrics presented in section 4.1, in particular the Sensitivity, specificity and BCR. These metrics are now defined for a specific time t, the time threshold at which the survival is computed.

The sensitivity/specificity is the proportion of high/low risk patients identified as such by a model. In the survival context, the sensitivity $SE(t)$ at time t is the proportion of high risk patients, *i.e.* classified in the high risk group, among those patients experiencing the event before time t. The specificity $SP(t)$ at time t is the proportion of low risk patients among those patients still at risk just after time t. The balanced classification rate $BCR(t)$ is the arithmetic average between sensitivity and specificity.

$$SE(t) = \frac{\sum_{i|t_i \leq t, \delta_i = 1} I(g_i = 1)}{\sum_{i|t_i \leq t, \delta_i = 1} 1} \tag{4.20}$$

$$SP(t) = \frac{\sum_{i|t_i > t} I(g_i = -1)}{\sum_{i|t_i > t} 1} \tag{4.21}$$

$$BCR(t) = \frac{SE(t) + SP(t)}{2} \tag{4.22}$$

$BCR(t)$ values lies between 0 and 1. A perfect prognostic index would have a $BCR(t)$ of 1. Uniform random guessing between risk groups has an expected $BCR(t)$ of 0.5 while a lower $BCR(t)$ would correspond to an even worse prognosis (*e.g.* inverting risk groups). It is worth noting that the patients censored before the time threshold are not assigned to any class. They are ignored in the BCR and the sensitivity.

4.2.6 Discussion

The general topic of **model performance assessment** is very broad and extensively discussed in the statistics and machine learning literature. It covers many aspects including several definitions of model performance scores or metrics, evaluation protocols (*e.g.* variants of cross-validation or bootstrap resampling), dedicated statistical tests to assess the significance of observed differences, to name just a few. The interested reader may consult, for example, a whole monograph on those issues, and many related ones, especially in a classification context [72].

For the sake of conciseness, our discussion in this section is much more focused. We are describing various performance metrics that are specific or adapted to survival data. Besides, while those metrics could in general be computed on the same (= training) data used to estimate a survival model, our interest is towards their use on independent **validation** data (either explicitly or through cross-validation or resampling). In other words, we are focusing throughout this work on **predictive performances** rather than **goodness-of-fit** measures. This is motivated, at least, by a pragmatic choice to report results, eventually to clinicians, that would be representative of what should be expected for new but similar patients[1]

Finally, we discuss here quality metrics that can be used to measure to which extent patients are assigned to a correct **risk group**. This

[1] Technically, this is related to an i.i.d. assumption between training and validation samples. This assumption is very common but could be questioned. Recent works on **transfer learning** [99] precisely aim at relaxing this assumption.

discretization corresponds to a decision support setting: "such decisions are typically binary and require the definition of clinically relevant decision thresholds" [128]. Considering a low versus a high risk group also simplifies the mathematical presentation of these notions. Yet, as mentioned below and in chapter 8, we also consider the cases with more than two risk groups.

While the classification metrics are mainly variations around the confusion matrix (table 4.1), the survival metrics are much more diverse. We cover here a non-exhaustive list of such metrics. The hazard ratio and C-index can assess both continuous risk scores and discrete risk groups. Others are designed solely for risk group assessment like the logrank, SEP and BCR. Our central question is: what should we use to assess the quality of the prediction into specific risk groups?

If we want to assess prediction into risk groups at a particular time threshold, the metrics of choice are the time dependent balanced classification rate, sensitivity and specificity. But these metrics are limited to a particular time threshold. They are unable to assess properly the survival through the complete follow-up.

In general, it is better to use the hazard ratio, C-index, logrank, SEP or D-index. To compare them, we use the set of desirable properties proposed by Royston and Sauerbrei [108] to define what is a good survival metric. Some of those properties are:

- **Interpretability**: The metric should have a simple and intuitive meaning.
- **Directness**: If the risk group ordering changes, the metric should change in the appropriate direction.
- **Unbiasedness**: The estimated metrics should also be an unbiased estimate of the true value. In particular, the metric should be close to a known value when the risk groups have on average no relationship with the survival times.

The logrank and SEP do not have the directness property, they are insensitive to the group ordering [28, 112]. For instance, in the special case of a high versus a low risk group, the two metrics are unable to detect when the groups are switched, for example by assigning all high risk patients to the low risk group. Moreover, these metrics do not penalize a situation where two out of k groups have exactly the same hazard. They cannot detect if there are more groups than what is needed. The logrank and SEP can only assess if there is a difference between the survival of some risk groups.

The C-index and hazard ratio have the directness property and are thus useful to quantify the difference between group survival [59]. But, this difference is useful and informative only if all groups incorporate a substantial part of the patients [108]. As they do not take into account the sizes of the groups, the C-index and hazard ratio can estimate the survival difference but not the relevance of risk groups. They are somehow biased as they do not tend to their minimal values when all patients tend to be in one group. The D-index also has the directness property and somehow takes into account the sizes of the groups. While being better than the C-index and hazard ratio, the D-index does not tend to its minimal value when all patients are in one group (see appendix C).

In chapter 8, we further investigate the differences between these survival metrics: hazard ratio, C-index, logrank, SEP and time dependent BCR. We also propose the balanced hazard ratio which have most desired properties of a good survival metric.

To sum up, we present and discuss in this chapter several performance metrics in the particular context of binary classification and survival risk group assessment. We limit our discussion essentially to predictive performance metrics for a **discrimination** purpose. The discrimination refers here to the capacity of a model to classify or to rank the patients in the right way. However, **calibration** is another aspect of prediction accuracy, which can be assessed separately [128, 75]. In particular, calibration measures the agreement between predictions and observed outcomes, *e.g.* if the estimated risk of a patient fits the observations and is neither overestimated nor underestimated. For example, the Hosmer-Lemeshow goodness-of-fit test [66] for the logistic model compares the predicted probabilities of event for each decile with the observed event rates. The calibration is particularly important for patients who are more concerned with their actual chances of survival rather than their relative risks. Another important aspect, which we do not further detail in this manuscript, is the estimation of the real impact of the predictive models on the patient **care** [104]. Indeed, the goal of predictive models is often to help physicians and to improve the clinical decision making. The impact of such decisions and whether they are ultimately beneficial or harmful to the patients should also be assessed.

Chapter 5

Feature Selection

A critical aspect of learning new prognostic models is the feature selection. Feature selection refers to the set of methods used to select a subset of relevant covariates/features, that can be used in a (prognostic) model. The feature selection can improved prognostic models in three key points: performance, interpretability and usability.

By selecting the relevant covariates, a good feature selection is expected to reduced the number of noisy ones. Eliminating these noisy variables can improve the prediction performance on unseen data. Feature selection can also be viewed as a form of regularization. It favours simpler models that are more likely to have better generalization performances. This increase in performance is not always observed, especially with drastic feature selection. An example is microarray where the number of features is sometimes divided by 1000, from 55000 to 50 genes/covariates. However, small losses in prediction performances are mitigated by the others advantages of the selection.

A major advantage of feature selection is interpretability. In microarray, by selecting a small number of relevant features, we have good chances to selected genes that are related to the disease under study. They can give us a better understanding of the biological processes involved. They can even be sometimes the target of new treatments.

By reducing the number of needed variables, we can also reduce the cost of a new prognostic model. The expression levels of the selected genes can be measured with some cheaper technologies than microarray. Technologies that might not measure thousands of genes but can be used to make small prognostic kits. Feature selection may make a prognostic model usable in clinics.

Classically, the feature selection schemes are divided in three cat-

egories [58]: filter, wrapper and embedded method. The **filters** are sometimes viewed as a preprocessing step. The selection is made separately of any prognostic model used later. They do not try to optimize directly the performance of the predictive model.

With the **wrappers**, the features are selected to maximize the predictive performance of the model. They often rely on iterative methods adding or removing features to search for the best ones. The models are here used as a black box to measure the quality of a set of variables. To avoid overfitting, the performances of these models should be estimated in cross-validation [58, 78]. This selection scheme requires the learning of a large number of models and is thus very time-consuming.

The feature selection in an **embedded method** is performed while learning the model. There is no separation between the feature selection and training. Many linear models can be turned in an embedded feature selection method by using some sparsity inducing regularization. An example of such regularization is the LASSO [132, 133] or elastic net [147].

In this thesis, we used filters and embedded methods. Filters are used for their computational efficiency. They are also less prone to overfitting and more stable [97, 58]. Stable in the sense that the selected features do not change much under small perturbations of the data. Two univariate filters are presented in the sections 5.1.1 and 5.1.2, respectively for classification and survival prediction.

The prognostic models used in this thesis are mainly linear ones such as the Cox and the logistic regression (see chapter 3). Such linear models are often regularized in high dimensions with an L_1 penalty and can be considered as embedded methods. The feature selection with L_1 type regularization is discussed in section 5.2.

In the context of gene expression data, such as microarray data, these sets of features/genes are defining gene signatures. They are called **gene signatures** as their combined expression pattern should be characteristic of a particular biological phenotype or medical condition. Section 5.3 presents some ways to assess the relevance of such gene signatures.

5.1 Univariate filters

Feature selection with filters is often associated with feature ranking. This ranking of the features is often done by computing their association with a response variable (group labels, survival times, *etc*).

We choose to present here two univariate filters (sections 5.1.1 and 5.1.2). They both rely on hypothesis testing to compute the association

between each feature and the response. They offer a ranking of the features and p-values associated to the tests.

To perform feature selection, there are two common choices:

- If the number d of features required is known, select the d first features with the smallest p-values.
- Otherwise, select the features with a p-value smaller than some threshold (often 0.05).

The second option requires the p-values to be corrected for the multiplicity of the test and *e.g.* to control the false discovery rate. In this thesis, we use the FDR procedure proposed by Benjamini and Hochberg [8].

5.1.1 T-test

The t-test is often used as a univariate filter for classification problems. For each feature, the t-test compare the mean values of the two risk groups.

$$t = \frac{|\mu_{\text{High}} - \mu_{\text{High}}|}{\sqrt{\sigma^2_{\text{High}}/n_{\text{High}} + \sigma^2_{\text{Low}}/n_{\text{Low}}}} \tag{5.1}$$

μ_{High} and μ_{Low} are the means of the feature in respectively the high and low risk groups. Those means are compared to the group variances and sizes, respectively: σ^2_{High}, σ^2_{Low}, n_{High} and n_{Low}. Under the null hypothesis that there is no difference between the groups, the statistic (5.1) follows a Student's t distribution from which we can compute a p-value.

5.1.2 Hazard ratio

Similarly as in section 4.2.1, an hazard ratio can be used to test the association between each feature and the survival. We replace here the risk scores and risk groups by a single covariate.

The hazard ratio of two patients i and k with respect to the feature j can be written:

$$\frac{h_i(t)}{h_k(t)} = \frac{h_0(t)\exp(\beta x_{ij})}{h_0(t)\exp(\beta x_{kj})} \tag{5.2}$$

$$= \frac{\exp(\beta x_{ij})}{\exp(\beta x_{kj})} = \exp(\beta(x_{ij} - x_{kj})) \tag{5.3}$$

The value $\exp(\beta)$ represents the hazard ratio between i and k, two patients with a unit difference $x_{ij} - x_{kj} = 1$ in feature j. As the gene expression is often expressed in $\log_2$, $\exp(\beta)$ is then the multiplicative increase of the hazard when the gene expression level is doubled. A p-value can be computed to test if β is significantly different from 0 using standard MLE inference (see section 2.4.1).

5.2 L_1 regularized embedded methods

Embedded feature selection methods differ from filter as the selection is performed while learning the model and not as a preprocessing step. The selected features are simply the feature that are used in model. Other features may be used during the learning but are not part of the final predictive model.

The L_1 regularization (or lasso) is a sparsity enforcing regularization [132] which is very often used in combination with (generalized) linear models such as the Cox or logistic regression. The lasso regularization can be viewed as some form of soft-thresholding of the model parameters. For instance, a L_1 regularized Cox model will tend to have its smallest parameters shrinked to zero. A parameter shrinked to zero means that the corresponding feature is not used in the model. The L_1 regularization can thus be used as multivariate feature selection.

The lasso regularization can however be too sparse, especially with correlated features [132, 147]. The lasso will tend to select one and discard all other correlated covariates. The elastic net regularization was introduced by Zou and Hastie [147] to solve this issue. It combines the lasso and the ridge (or L_2) penalties.

$$L_e = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 \tag{5.4}$$

$$= \left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2\right) \tag{5.5}$$

where $\alpha \in [0-1]$ is a meta-parameter of the elastic-net regularization. It combines the advantage of both regularization. The elastic-net is sparse and can be used as feature selection. It also encourages the grouping effect. Highly correlated features tend to be in or out of the model together [147].

5.3 Gene signature validation

Assessing the relevance of gene signatures is most of the time an unsupervised problem. Excepted with artificial experiments, the best set of features is unknown. There are two common indirect ways to assess the relevance of a gene signature.

The first possibility is to assess its predictive performance, especially when the selected features are used to build a prognostic model. The validation of a gene signature will then depend on the choice of model and performance metric (chapters 3 and 4). This assessment must be done on independent data, not used for the feature selection, to avoid any optimistic bias.

The obtained performances could be compared with other prognostic models or gene signatures on the same data. In particular, the gene signature could be compared to random gene signatures. This approach was proposed by Venet *et al.* [142] and is described in section 5.3.1.

The second possibility comes from the idea that a good feature selection should select genes related to the disease under study. In particular, the signature is expected to have links with other known signatures. For instance, signatures of metabolic or signaling pathways that are related to the predictive task. Rediscovering those links can be seen as a validation of the biological soundness of the signature.

Many tools are available to compare gene signatures [67]. Those tools, known as singular enrichment analysis, commonly relies on an hypergeometric or Fisher's exact test. Section 5.3.2 presents the principles of this enrichment analysis.

The feature selection can often be seen as the selection of a cutoff on a particular ranking of the features (see section 5.1). Subramanian *et al.* [129] proposed to compare directly this ranking to other reference gene signature to avoid any cutoff effect. This method known as gene set enrichment analysis (GSEA) is presented in section 5.3.3.

A third possibility, out of the scope of this thesis, is a true biological validation of the genes in a lab.

5.3.1 Predictive performance of random gene signatures

Venet *et al.* [142] show that most random gene signatures are significantly associated with breast cancer outcome. They report more than 90% of the signatures with > 100 genes tested as significant.

Since there are many genes associated with cancer progression, even random signatures can contain useful genes. As a practical example, assuming that there are only 1000 probesets associated with cancer

progression on a total of 22283 probesets in an Affymetrix HGU133a microarray, the probability of a random signature of 100 probesets to contain at least one of them is 0.99.

To validate the interest of a signature, this signature should be significantly better at the predictive task than random signatures of the same size. In particular the signature should be better than 95% of the random signatures (using the statistical standard of p-values < 0.05).

This methodology was used to validated the prognostic performances of the CycHyp and ContHyp signatures on breast cancer data (presented in section 10). Using the same experimental protocol, we compare the logrank p-values (see section 4.2.3) of the CycHyp and ContHyp signatures with random signature of the same sizes (resp. 87 and 123 probe sets). Figure 5.1 shows the distribution of the p-values (logrank test in log 10) for 1000 randomly generated signatures together with the p-values of the CycHyp and ContHyp signatures represented with the two red dots. The discrimination between risk groups was here significantly higher (p-value < 0.001) with the CycHyp signature as compared to each of the random signatures whereas the ContHyp signature was not significantly better (*vs.* random ones; p-value=0.141). The green line in figure 5.1 represent the 0.05 threshold for the p-values. We can observe that logrank p-values of most random gene signatures are below this threshold and are thus significant.

5.3.2 Singular enrichment analysis

The traditional strategy of singular enrichment analysis is to compare the gene signature with a set of reference signatures/genesets. An example of application is TFactS [43]. In TFactS, the gene signature is compared to genesets containing target genes of transcription factors. This tool can thus test the association between a particular signature and the regulation of transcription factors.

To assess the links between a signature and a reference geneset, tools like TFacts compare their number of genes in common with respect to what is expected from a random selection. The comparison between two signatures can be summarized in a table:

		Gene Signature		
		Present	Absent	Total
Reference	Present	k	$l-k$	l
	Absent	$d-k$	$p+k-l-d$	$p-l$
	Total	d	$p-d$	p

Table 5.1: Comparison between gene signatures

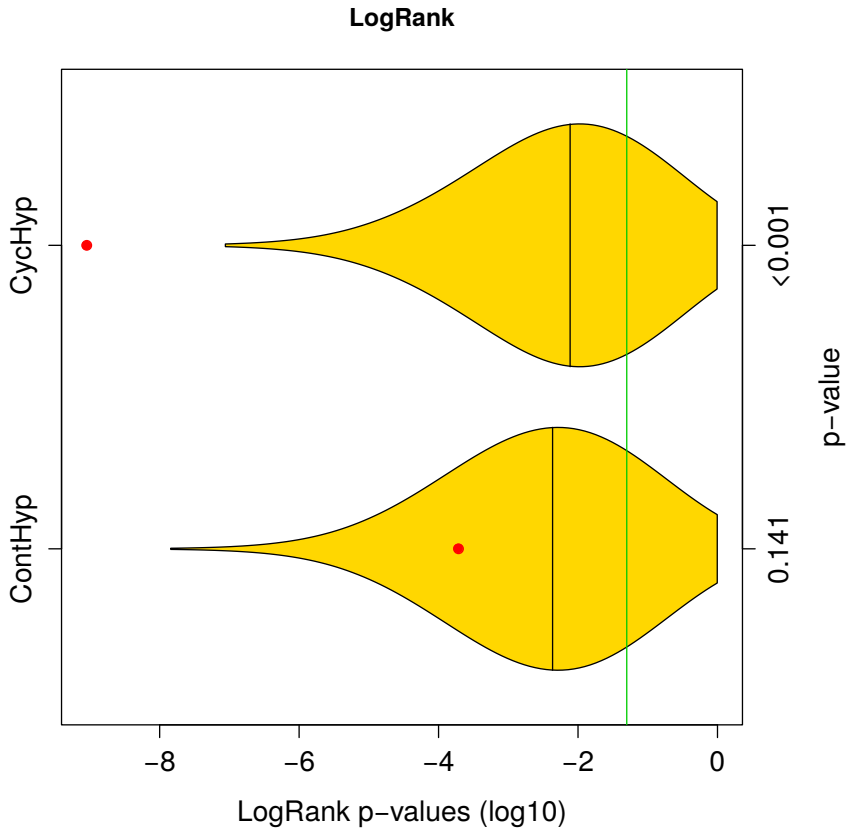


Figure 5.1: Graph represents the power of discrimination in high *vs.* low risk groups (expressed as the logarithm of the p-values of the logrank) of the ContHyp and CycHyp signatures (see red dots) versus 1,000 randomly generated signatures (yellow shapes depicting their distribution).

where d is the number of genes selected in the signature on a total of p genes available, l is the number of genes in the reference signature, k is the number of genes in common between the two signatures.

Under the null hypothesis that there is no link between the two signatures, k follows a hypergeometric distribution. The probability to observe k or more genes in common under this null hypothesis gives the p-values of the enrichment.

Singular enrichment analysis has the problem of being sensitive to the number of genes selected in the signature. If this number is too small, the analysis may miss some important genesets. At the opposite, selecting too many genes may reduce the significance of the test.

5.3.3 Gene set enrichment analysis

Mootha *et al.* [96] introduce the gene set enrichment analysis to avoid the cutoff effect of the traditional singular enrichment analysis. The principle of gene set enrichment analysis is to test if there is an association between a ranking of the feature, as proposed for example with a filter (see section 5.1), and reference gene signatures. In particular, we test if the genes of a reference signature are randomly distributed in the ranking of the features.

For example, we could have ranked the genes according to an univariate hazard-ratio. A gene set associated with the survival of patients should have its genes at the top or at the bottom of list. Genes at the top (resp. bottom) are correlated with shorter (resp. longer) survival.

For a gene signature of size l, the enrichment score (ES) as proposed by Mootha *et al.* [96] can be computed from a ranked list of all the p genes. Walking down this list, we increase or decrease a running-sum statistic by some value w_j for each gene j encountered.

$$w_j = \begin{cases} \sqrt{\frac{(p-l)}{l}} & \text{if} \quad j \text{ is in the signature} \\ -\sqrt{\frac{l}{(p-l)}} & \text{otherwise} \end{cases} \tag{5.6}$$

In particular, the running-sum statistic is define as 0 at the beginning and the end of the list:

$$\sum_{j=1}^{0} w_j = 0 \tag{5.7}$$

$$\sum_{j=1}^{p} w_j = \sqrt{\frac{(p-l)}{l}} l - \sqrt{\frac{l}{(p-l)}} (p-l) = 0 \tag{5.8}$$

When a geneset is randomly distributed in a ranked list of genes, the positive and negative w_j compensate each others and the running-sum should stay close to zero. Its maximum, the enrichment score, is computed as:

$$\mathrm{ES} = \max_{1 \leq k \leq p} \sum_{j=1}^{k} w_j \tag{5.9}$$

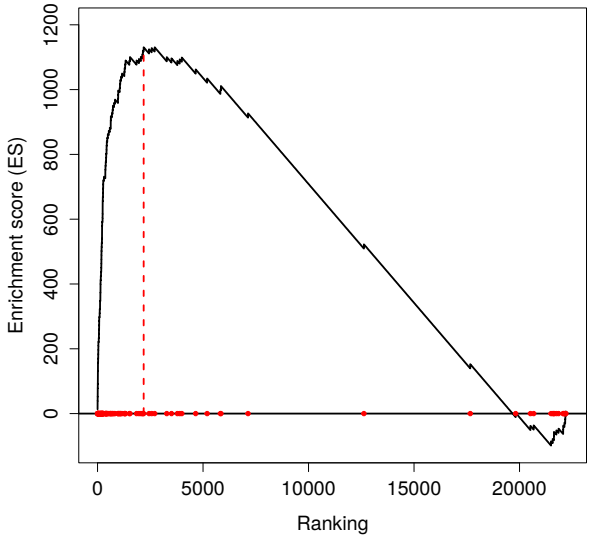
The enrichment score (ES) is reported and can be compared with a null distribution. The null distribution is estimated while computing the enrichment score on permutations of the class labels or survival times.

As an example, we compute the enrichment score (ES) of two signatures on breast cancer data. The ranking of the features is based on their correlation with the tumor grade. The first signature is the GGI signature [124] which was designed precisely to be predictive of the histological grade. The second signature is the Gene76 signature [143] and is predictive of the breast cancer outcome. Figure 5.2 presents the values of the running-sum statistic (black curves). The red dots represent the positions of the signatures (one for each gene) in the ranking. The dashed red lines represent the position of the running-sum statistic maximums (where the enrichment score is computed). The GGI signature, in constrast with Gene76, is mostly enriched at the beginning of the ranking and has a higher enrichment score, as expected.

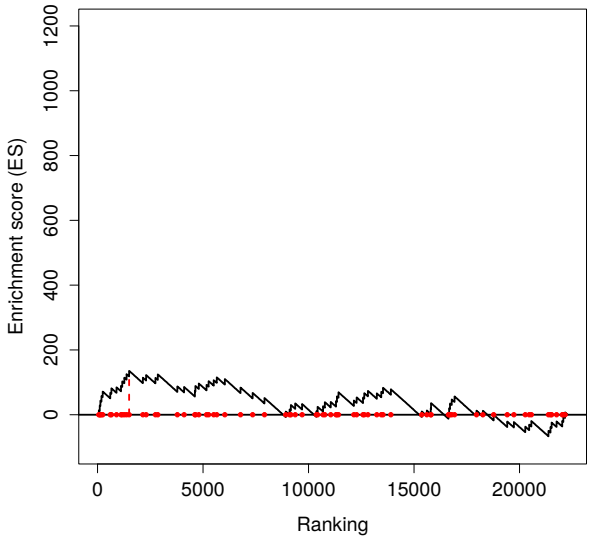
5.3.4 Discussion

The validation of a gene signature is a very complicated process where a particular care should be taken to avoid any optimistic bias. This bias can come from the protocol used for the feature selection and the validation [4]. It can also come from the data, indeed it is quite common to have patients shared between different datasets. This validation also requires a proper performance assessment with the right metrics (see chapters 4 and 8).

The validation of a signature with enrichment analysis is not perfect either. First, we depend on the set of published gene signatures that are assumed complete and well annotated. Finding significantly enriched related genesets can be used as a biological validation of the signature. But, finding no significantly enriched genesets is difficult to interpret and cannot be used to reject a signature. Another question is the overlap between these genesets and their interactions, enrichment analysis are still essentially univariate [67].



(a) GGI signature



(b) Gene76 signature

Figure 5.2: Enrichment score of the GGI [124] and Gene76 [143] signatures. The ranking of the probesets is based on the correlation with the tumor grades on a breast cancer dataset.

Moreover, singular enrichment analysis (section 5.3.2) is sensitive to the size of the signature. If this number is too small, the analysis may miss some important genesets. At the opposite, selecting too many genes may reduce the significance of the test. Gene set enrichment analysis (section 5.3.3) try to solve this issue while assessing a ranking and not the true signature. GSEA can thus be problematic for feature selections without ranking such as the embedded methods presented in section 5.2.

Part II

Methodological Contributions

Chapter 6

Coxlogit Model

6.1 Introduction

As discussed in chapter 3, predicting risk groups can be seen as a classification or a survival prediction problem. The difference between the two approaches is the kind of supervision available. The Cox model, presented in section 2.4 and 3.3.1, is the most commonly used with survival data (t_i, δ_i). When the risk groups $y_i \in \{-1, 1\}$ are known, classification models such as SVMs or logistic models are preferred (see section 3.2).

Sometimes, these two kinds of supervision are available together for the same data. In breast cancer for example, risk groups can be defined from the tumor grades or from the survival data. A logistic model predicting the grades might be very close to a Cox model predicting the survival as both tasks are highly correlated. The originality of this contribution is to tackle both problems jointly.

We consider in particular generalized linear models as they offer a direct interpretation in terms of individual feature relevances. The proposed Coxlogit model is a natural extension to logistic regression for which we assume that the survival times and class labels are random variables conditioned by a common risk. We show that the partial likelihood of such model, to fit the ordering of observed survival times, is directly related to the logistic class probabilities. Learning can then be expressed as maximizing the joint probability of class labels and the ordering of survival events, conditioned to a common weight vector.

Embedded feature selection follows naturally when fitting such a model with a LASSO or elastic net penalty. Such penalties prevent overfitting while enforcing a common sparse support. Learning is also a convex problem that can be efficiently solved through a coordinate

descent algorithm.

- Branders, S., D'Ambrosio, R., and Dupont, P. (2014). The Coxlogit model: Feature selection from survival and classification data. In *2014 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making (MCDM)*, pages 137–143
- Branders, S., D'Ambrosio, R., and Dupont, P. (2015a). A mixture Cox-Logistic model for feature selection from survival and classification data. *arXiv:1502.01493 [stat.ML]*, pages 1–6

6.2 The Coxlogit approach

One considers a survival analysis framework made of a collection of samples and their associated survival times, which are possibly censored. One further assumes that each training sample is labeled into a specific risk group. Formally, each sample $i \in \{1, \ldots, n\}$ is then characterized by a 4-tuple $(t_i, \delta_i, y_i, \mathbf{x}_i)$ with the survival data, the binary class label and the patient covariates.

The survival data and the class label of patient i are seen here as two observations of random variables conditioned by a common risk of event, r_i. This risk is simply modeled as a linear combination of the sample covariates ($\mathbf{x}_i \in \mathbb{R}^p$) : $r_i = \boldsymbol{\beta}^\top \mathbf{x}_i$ but the fit of the parameters $\boldsymbol{\beta}$ should consider both types of supervision.

Starting from the classification viewpoint, a logistic regression predicts from the vector $\mathbf{x}_i$ the probability of patient i to be in a specific group:

$$P(Y_i = 1|\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{6.1}$$

$$P(Y_i = -1|\mathbf{x}_i) = \frac{1}{1 + \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{6.2}$$

$$= 1 - P(Y_i = 1|\mathbf{x}_i) \tag{6.3}$$

The risk score of a patient, $r_i = \boldsymbol{\beta}^\top \mathbf{x}_i$, can be interpreted through the logistic model as class probabilities: high risk patients are more likely to be in the high risk group +1 and a zero risk score corresponds to an equal probability to be in either risk groups. The likelihood of the parameters $\boldsymbol{\beta}$ with respect to the observed labels y_i is given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y_i = y_i|\mathbf{x}_i) \tag{6.4}$$

$$= \prod_{i=1}^{n} \frac{1}{1 + \exp(-y_i(\boldsymbol{\beta}^\top \mathbf{x}_i))} \tag{6.5}$$

Looking now at the survival times and knowing that an event occurs at time t_i, one typically computes the probability of patient i having the event over the set of patients still at risk just before time t_i : $R(t_i) = \{k|t_k \geq t_i\}$.

Since high risk patients tend to have the event before low risk ones, the likelihood of this observed event can be modelled as the conditional probability of patient i being the only high risk patient, knowing that exactly one patient is in the high risk group. The conditional probability is computed as the probability of patient i being in the high risk group and all others ($R(t_i)\backslash\{i\}$) being in the low risk group divided by the probability of having only one patient in the high risk group. This likelihood can be expressed in terms of the logistic class probabilities $P(Y_i = 1|\mathbf{x}_i)$ and $P(Y_i = -1|\mathbf{x}_i)$:

$$L_i(\boldsymbol{\beta}) = \frac{P(Y_i = 1|\mathbf{x}_i) \prod_{j \in R(t_i)\backslash\{i\}} P(Y_j = -1|\mathbf{x}_j)}{\sum_{k \in R(t_i)} P(Y_k = 1|\mathbf{x}_k) \prod_{j \in R(t_i)\backslash\{k\}} P(Y_j = -1|\mathbf{x}_j)} \tag{6.6}$$

$$= \frac{\frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)} \prod_{j \in R(t_i)\backslash\{i\}} \frac{1}{1+\exp(\boldsymbol{\beta}^\top \mathbf{x}_j)}}{\sum_{k \in R(t_i)} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_k)}{1+\exp(\boldsymbol{\beta}^\top \mathbf{x}_k)} \prod_{j \in R(t_i)\backslash\{k\}} \frac{1}{1+\exp(\boldsymbol{\beta}^\top \mathbf{x}_j)}} \tag{6.7}$$

$$= \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{k \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k)} \tag{6.8}$$

Expression (6.8), aggregated over all events, boils down to the partial likelihood (section 2.4.1) of a Cox model for survival data. The relation between the conditional logistic model [27, 28] and the cox model is used here in the Coxlogit model to link the class probabilities and survival times.

The likelihood of the Coxlogit model is now defined as the joint probability of the observed events and risk group labels knowing the parameters $\boldsymbol{\beta}$. Assuming the labels and the times to event to be conditionally independent *given* those parameters, this likelihood can be computed as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{1+\exp(-y_i(\boldsymbol{\beta}^\top \mathbf{x}_i))} \left[\frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{k \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k)} \right]^{\delta_i} \tag{6.9}$$

The loss of $\mathcal{L}(\boldsymbol{\beta})$ of the Coxlogit model is thus naturally formulated as a mixture of a Cox and logistic losses:

$$\mathcal{L}(\boldsymbol{\beta}) = \mathcal{L}_{\text{cox}}(\boldsymbol{\beta}) + \mathcal{L}_{\text{logi}}(\boldsymbol{\beta}) \tag{6.10}$$

In the derivation of the Coxlogit model, we use a similar interpretation of the Cox model as the one presented in section 3.3.1. Compared with the Cox model, the Coxlogit model can take into account the prior knowledge available on the risk groups, *i.e.* the risk group labels y_i. These labels combined with the logistic loss allow us to interpret the risk scores $r_i = \boldsymbol{\beta}^\top \mathbf{x}_i$ in terms of risk group probabilities. Unlike the Cox model, the Coxlogit risk scores can directly be used in a decision function:

$$g_i = \begin{cases} 1 & \text{if} \quad P(Y_i = 1|\mathbf{x}_i) > 0.5 \\ -1 & \text{if} \quad P(Y_i = -1|\mathbf{x}_i) > 0.5 \end{cases} \tag{6.11}$$

$$= \text{sign}(\boldsymbol{\beta}^\top \mathbf{x}_i) \tag{6.12}$$

From equation (6.8), we can observed that the Cox partial likelihood is invariant by translation, *i.e.* $\boldsymbol{\beta}^\top \mathbf{x}_i$ is equivalent to $\boldsymbol{\beta}^\top \mathbf{x}_i + \theta$. This is not the case of the logistic likelihood, equation (6.5). In the Coxlogit model, the risk scores are thus learned from both the labels and the survival times but the decision function is mainly based on the class labels.

6.2.1 The Coxlogit mixed model

In equation (6.10), we define the Coxlogit loss as the sum of the Cox and logistic losses. For a better control on the model, we can further define the Coxlogit model as a mixture of both models:

$$\mathcal{L}(\boldsymbol{\beta}) = (1-\gamma)\ \mathcal{L}_{\text{cox}}(\boldsymbol{\beta}) + \ \gamma\ \mathcal{L}_{\text{logi}}(\boldsymbol{\beta}) \tag{6.13}$$

The meta-parameter $\gamma \in [0-1]$ controls the contribution of either losses in the model, with $\gamma = 0$ (respectively $\gamma = 1$) corresponding to a pure Cox model (respectively a pure logistic regression). When γ is not

explicitly specified, the Coxlogit model refers to the model presented in section 6.2, or equivalently to a Coxlogit model with a γ of 0.5.

We use this definition (6.13) of the Coxlogit model with γ to have a better understanding of the roles of each loss in the model.

6.2.2 Feature selection with the Coxlogit model

An embedded feature selection is performed by regularizing the objective function of the Coxlogit model (6.10):

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\beta}) + \lambda \mathcal{R}(\boldsymbol{\beta}) \tag{6.14}$$

where $\mathcal{R}(\boldsymbol{\beta})$ is a sparsity enforcing regularization such as LASSO [132, 133] or elastic net [147], and $\lambda > 0$ a regularization constant. A coordinate descent algorithm, adapted from [52, 117], is used here to solve this convex problem. It starts from a trivial solution ($\boldsymbol{\beta} = 0$) for a large λ, and follows the regularization path when λ is gradually decreased till the model includes a desired number of features (= non-zero weight values). The algorithm we implemented is described in more details in appendix B.

6.3 Experiments

In order to validate the proposed Coxlogit model in prediction and feature selection tasks, we perform experiments on synthetic and real datasets. The performances in classification and survival prediction are assessed according to the classification accuracy (see section 4.1.1) and the risk group concordance index (see section 4.12), respectively.

In this chapter, the results are reported in accuracy as all classification problems are almost perfectly balanced. In such cases, the classification accuracy is equivalent to the balanced classification rate (see section 4.1.3).

The global performance, while takling both tasks, is computed as the harmonic average between classification accuracy and C-index:

$$H_{\text{mean}} = \left(\frac{\text{Accuracy}^{-1} + \text{C-Index}^{-1}}{2} \right)^{-1} \tag{6.15}$$

We use the harmonic average to favor solutions with both a high accuracy and a high C-index. The significance of differences between

models is computed according to a Friedman test with a Nemenyi post-hoc test [34].

The first experiments reported on a synthetic dataset (section 6.3.1) illustrate that the Coxlogit model is able to select features that are jointly informative for survival and subgroup classification. Even without a perfect correlation between the two tasks, the Coxlogit approach offers competitive classification and survival prediction results. Those results are confirmed on real breast cancer prognosis studies (section 6.3.3).

6.3.1 Synthetic data

The synthetic dataset is designed to have both supervisions in terms of survival times and subgroup classification. The dataset is designed with four groups of features. The first 3 groups of features are predictive of respectively

- both the survival and the class label.
- the survival only.
- the class label only.

while the remaining features are purely random and supposed to represent noise.

One would like to assess to which extent the Coxlogit approach is able to predict both the class label and survival, as compared to a regularized Cox or logistic model alone. One could also compare the features selected by the three models. The data matrix $X \in \mathbb{R}^{n \times p}$ is drawn from a $\mathcal{N}(0, 1)$ distribution to represent covariates that have been centered and normalized to unit variance, a common practice in our context.

The class assignments and hazards are generated from distinct linear combinations of the features. The weights of those predictors, β_j, are drawn from a uniform distribution over $[-1, -0.5] \cup [0.5, 1]$. Class labels $y \in \{-1, 1\}^n$ are generated from the first and third groups in the following way:

$$\boldsymbol{\beta}_{class} = \big(\beta_j \quad 0 \dots 0 \quad \beta_j \quad 0 \dots 0 \big)$$
$$y_i = \operatorname{sign}\left(\boldsymbol{\beta}_{class}^{\top} \mathbf{x}_i\right)$$

The survival data (t_i, δ_i) are generated from two Weibull distributions, for the time to event and the censoring respectively. The Weibull distribution for the time to event is parametrized such that the hazard

$h_i(t)$ depends on the features from the first and second groups :

$$\boldsymbol{\beta}_{surv} = (\ \beta_j \quad \beta_j \quad 0\dots0 \quad 0\dots0\ )$$
$$h_i(t) \propto \exp(\boldsymbol{\beta}_{surv}^\top \mathbf{x}_i)$$

The class labels and survival times are correlated as $\boldsymbol{\beta}_{class}$ and $\boldsymbol{\beta}_{surv}$ have the same weights on the first groups of features. To increase or decrease the correlation between the two supervision, we change the sizes of the second and third groups containing features respectively predictive of the survival or the classification. The smaller is their size, the higher is the correlation between class labels and survival times.

In practice, we consider a dataset of $n = 300$ patients and $p = 100$ features. 100 samples are used for training and 200 independent samples serve as validation. The number of features in the first group, predictive of both the survival and the class label, is fixed to 10. The sizes of the second and third groups change from $k = 2$ to $k = 10$.

6.3.2 Results on synthetic data

This section reports the results using the Coxlogit model on the synthetic data described above, as compared to a regularized Cox model ($\gamma = 0$) or a logistic regression ($\gamma = 1$). For each experiment, the regularization path is followed till the model contains exactly 10 features. The absolute weight value assigned to each feature can be easily interpreted as the relevance of the features estimated by the model.

Figure 6.1 reports the model weights obtained while varying γ in $[0, 1]$. The data are generated here with 10 features in each of the three groups. Figure 6.1 shows a smooth transition between the features selected by the method while varying the value of γ. The Cox model only selects features that contains some survival informations (in green and red). Similarly, the regularized logistic model only selects features associated to class assignments (in green and blue), plus one random feature in this particular run. In contrast, the Coxlogit model (typically for $\gamma = 0.5$), tends to select only those features (in green) that are informative for both tasks.

Figure 6.2 reports the sum of the absolute weight values in each group while repeating the above experiment 100 times and averaging those absolute values over the 100 runs. While the Cox model, respectively the logistic regression, always selects features related to the survival, respectively the classification in subgroups, the Coxlogit model clearly favors the selection of common features.

Figures 6.3, 6.4 and 6.5 report the predictive results, respectively in terms of classification accuracy, C-index and harmonic average between

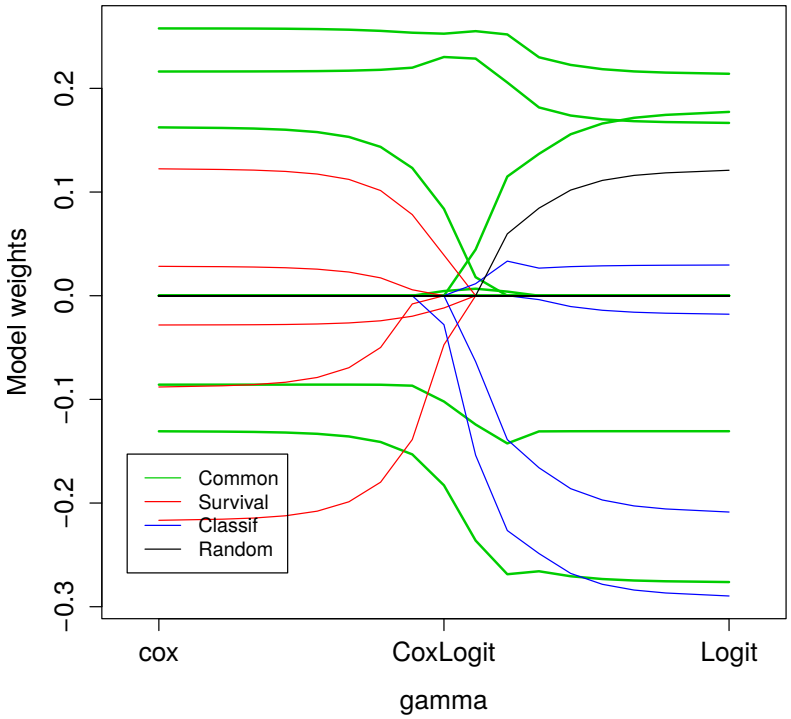


Figure 6.1: Model weights obtained while varying γ from 0 (Cox model) to 1 (Logistic model). The absolute value of the weights represent the importance assigned by the model to each feature. Only 10 out of the 100 weights have a non-zero value as a consequence of the chosen working condition along the regularization path.

them, obtained using the features selected by the Coxlogit model and averaged over 100 runs.

The results presented in figure 6.3 show that the best classification accuracy is obtained for a logistic regression or a Coxlogit model with $\gamma \geq 0.75$. Logically, a model fitted to maximize only a regularized Cox log-likelihood ($\gamma = 0$) performs poorly according to this metric.

Similarly, as reported in figure 6.4, the best C-index is obtained for a Cox model or a Coxlogit model with $\gamma \leq 0.25$ while a model fitted to maximize only a regularized logistic log-likelihood ($\gamma = 1$) is poor at predicting survival times.

According to a Friedman test with a Nemenyi post-hoc test [34], the Coxlogit model ($\gamma = 0.5$) is not significantly different from the best performances in classification (accuracy) or survival prediction (C-index). This result is confirmed in figure 6.5 reporting the harmonic average between C-index and classification accuracy. The best results are obtained

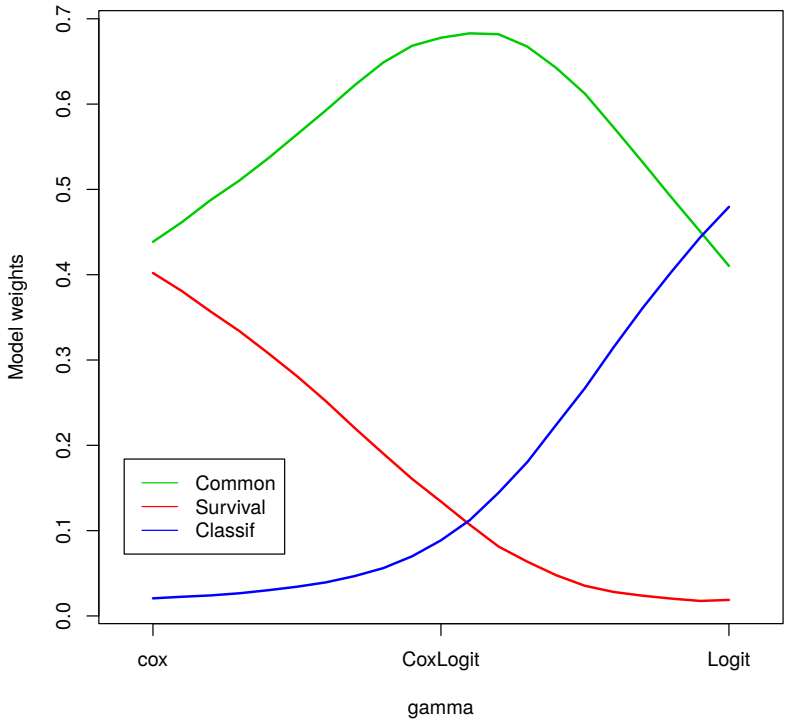


Figure 6.2: Mean absolute weight values in each group of features (Common, Survival, Classification, Random) computed while varying γ within $[0, 1]$.

with the Coxlogit model ($\gamma = 0.5$) which is here significantly better than either a Cox or a logistic regression model.

These results are obtained on data generated with 10 relevant features in each of the three groups. The two supervisions, class labels and survival data, are thus not exactly two different supervisions of unique problem. The class labels and survival data are generated here from two different true risk scores: $\boldsymbol{\beta}_{class}^{\top}\mathbf{x}_i$ and $\boldsymbol{\beta}_{surv}^{\top}\mathbf{x}_i$. To estimate how close are the classification and survival prediction tasks, we compute the Pearson correlation coefficient between these true risk scores. In this experiment, the correlation is on average equal to 0.5. The Coxlogit model will then have to find a compromise and cannot be perfect. However, the Coxlogit model is able here to perform comparably at both tasks as compared to a Cox or a logistic model.

To strengthen the links between the class labels and survival times, we generate data with 10 common, 2 survival and 2 classification features. The Pearson correlation coefficient between $\boldsymbol{\beta}_{class}^{\top}\mathbf{x}_i$ and $\boldsymbol{\beta}_{surv}^{\top}\mathbf{x}_i$ is on average equal to 0.83 in this setting. Figure 6.6 reports the pre-

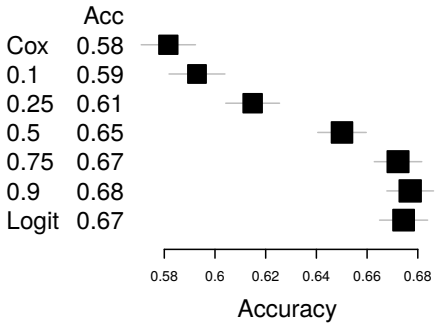


Figure 6.3: Forest plot of the average classification accuracy obtained while varying γ from 0 (Cox model) to 1 (Logistic model). These results are averaged on 100 repeated experiments with synthetic datasets generated with 10 common, 10 survival and 10 classification features.

diction results on these new data. The Coxlogit model ($\gamma = 0.5$) is significantly better in C-index and harmonic mean than a logistic regression or a Cox model. The Coxlogit model makes here a better use of both surpervisions to improve the performances in classification and survival prediction.

Those results can be explain by thinking of each supervision as an additional regularization. The Cox part of the Coxlogit model prevents an overfitting on the class labels, and *vice versa*.

The improvement can also be explain by a better feature selection. Figure 6.7 reports the average number of relevant features selected by the models. The relevant features are defined as the features used to generate the class labels or the survival data, *i.e.* the 10 common, the 2 survival and the 2 classification features. On average, the Coxlogit model ($\gamma = 0.5$) finds more than 8 relevant features on the 10 selected which is significantly more than either a Cox or a logistic regression model.

To further test the feature selection, we generate datasets with $p = 1000$ features: 10 common, 0 survival, 0 classification and 990 random features. With no survival or classification features, $\boldsymbol{\beta}_{class}^\top \mathbf{x}_i$ becomes equal to $\boldsymbol{\beta}_{surv}^\top \mathbf{x}_i$. The survival data and the class labels are thus generated from the same features and same risk scores, which allows us to have a fair comparison between the different models.

Figure 6.8 reports the average number of relevant features selected. The feature selection of Coxlogit model ($\gamma = 0.5$) is significantly better than either a Cox or a logistic regression model.

These results in feature selection are confirmed by the predictive

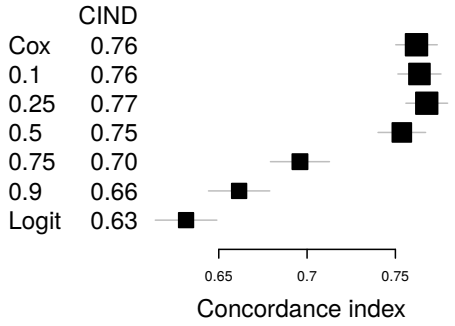


Figure 6.4: Forest plot of the average C-index obtained while varying γ from 0 (Cox model) to 1 (Logistic model). These results are averaged on 100 repeated experiments with synthetic datasets generated with 10 common, 10 survival and 10 classification features.

performances in figure 6.9. The Coxlogit model is significantly better in terms of classification accuracy, C-index and harmonic average between them.

The previous results show that the Coxlogit model is better when the two tasks are strongly connected. To assess the behaviour of the model when there is no link between the tasks, we generate datasets with $p = 100$ features: 0 common, 10 survival, 10 classification and 80 random features. Figure 6.10 reports the predictive performances of the models on these data sets. The best model for the classification (resp. survival prediction) is the logistic model (resp. Cox model). These two models are also the worst when looking at the other task (Cox in classification and logistic model in survival prediction). The Coxlogit model is a trade-off between the two with average results. With no link between the tasks, the Coxlogit model has no benefit of using both supervisions. However looking at their harmonic average, the Coxlogit model ($\gamma = 0.5$) is still significantly better than the Cox and logistic model.

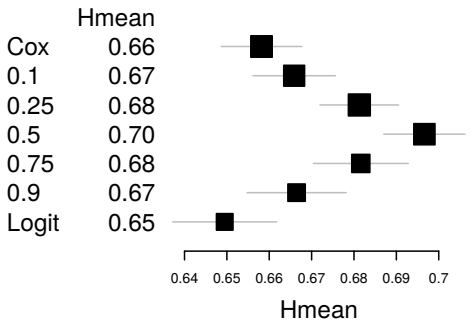


Figure 6.5: Forest plot with the harmonic average between classification accuracy and C-index obtained while varying γ from 0 (Cox model) to 1 (Logistic model). These results are averaged on 100 repeated experiments with synthetic datasets generated with 10 common, 10 survival and 10 classification features.

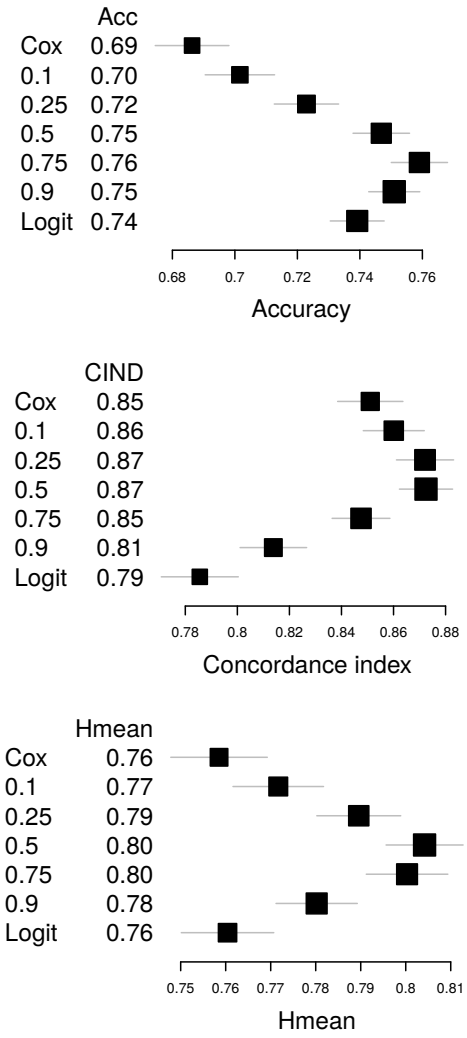


Figure 6.6: Forest plot with the results in accuracy, C-index and their harmonic average obtained while varying γ from 0 (Cox model) to 1 (Logistic model). These results are averaged on 100 repeated experiments with synthetic datasets generated with a high correlation between class labels and survival times: 10 common, 2 survival and 2 classification features.

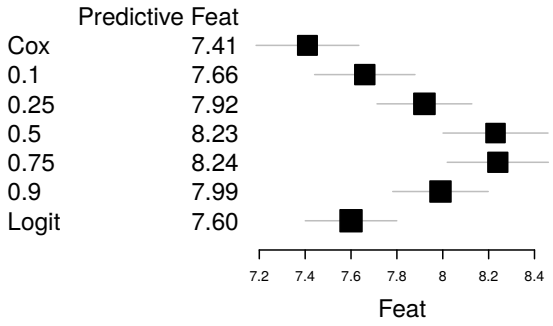


Figure 6.7: Forest plot with the average number of relevant features amongst the 10 selected in each model. Each results is an average on 100 repeated experiments with synthetic datasets generated with a high correlation between class labels and survival times: 10 common, 2 survival and 2 classification features.

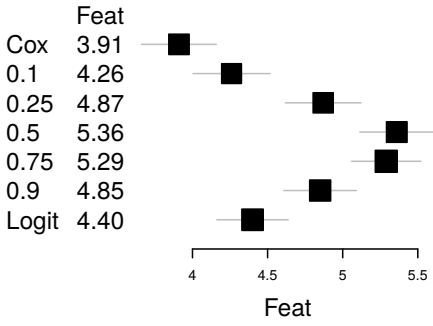


Figure 6.8: Forest plot with the average number of relevant features amongst the 10 selected in each model. Each results is an average on 100 repeated experiments with synthetic datasets. The data are generated with a perfect correlation between class labels and survival times (with no survival nor classification features).

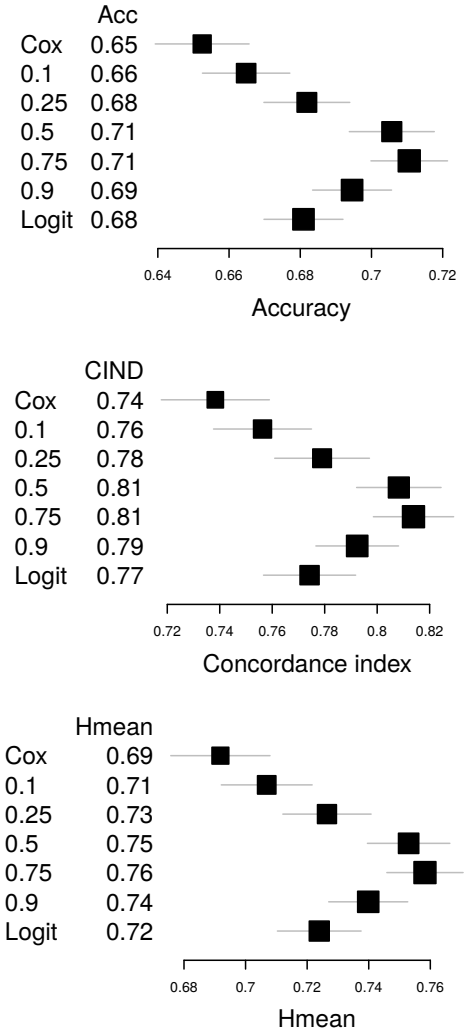


Figure 6.9: Forest plot with the results in accuracy, C-index and their harmonic average obtained while varying γ from 0 (Cox model) to 1 (Logistic model). Each results is an average on 100 repeated experiments with synthetic datasets. The data are generated with a perfect correlation between class labels and survival times (with no survival nor classification features).

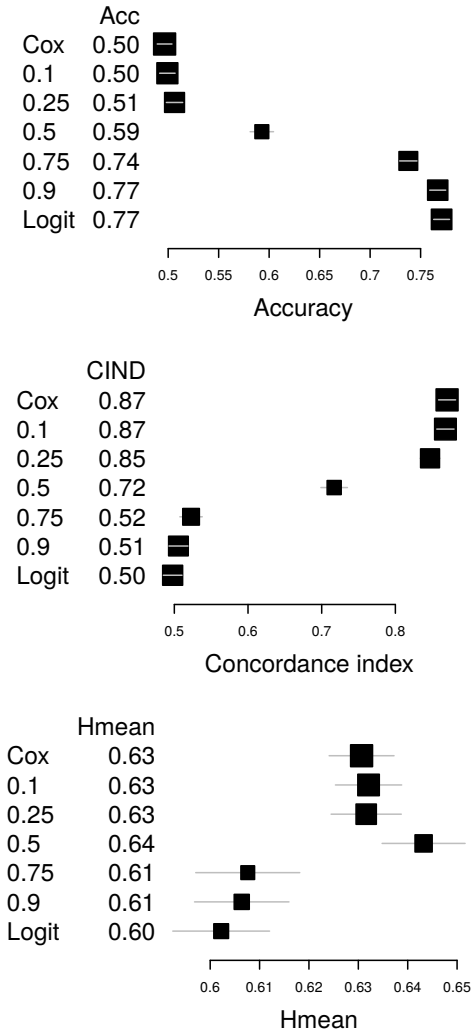


Figure 6.10: Forest plot with the results in accuracy, C-index and their harmonic average obtained while varying γ from 0 (Cox model) to 1 (Logistic model). Each results is an average on 100 repeated experiments with synthetic datasets. The data are generated with no links between class labels and survival times (with no common features).

6.3.3 Results on breast cancer data

We further assess the Coxlogit approach on 4 breast cancer studies (GSE2034, GSE5327, GSE7390, GSE2990) from the GEO repository. Those samples are gene expression values measured on the Affymetrix HGU133a microarray platform and distant metastasis is used as survival end point. All samples are gathered in a common dataset including 554 patients and 1115 features, after keeping only the dimensions with the largest variances. The objective is to predict both the grade of the tumor [54], discretized into low versus high grade with roughly equal priors, and the survival probability of the patients.

The 98 and 173 patients with respectively a tumor grade of 1 (well differentiated) and 2 (moderately differentiated) are grouped in the low risk groups, $y_i = -1$. The 283 patients with grade 3 (poorly differentiated) are the high risk patients, $y_i = 1$. To measure the differences between these two tasks, we compute the C-index of the class labels which is equal to 0.7. This confirms that the tumor grade is associated with the survival of patients.

Results are reported below over 100 resamplings without replacement (also known as subsampling), the training set size varies from 20 to 400 patients, the test set contains the 154 other patients. In the first experiment, 10 genes are selected by the three methods. Figure 6.11 present the results in accuracy, C-index and the harmonic mean between both.

The logistic and Coxlogit model have comparable results in accuracy and are significantly better than the cox model. In survival prediction, logistic and Coxlogit model are better than the Cox model for the training sets with less than 200 patients. As expected, the maximum C-index of the logistic model is 0.7 which is the C-index of the patient tumor grades on these data. Using both supervision, the Coxlogit model is able here to improve the survival prediction of the logistic and cox model. Figures 6.12 and 6.13 present similar results with 2, 20 and 50 genes selected.

6.4 Conclusion and perspectives

Classification and survival prediction are two common tasks in cancer research. These two tasks are often two different views of a unique problem of predicting risk groups. With the Coxlogit model, we propose to model together these two supervisions: class labels y_i and survival data (t_i, δ_i).

Using both supervisions introduces more constraints in the Coxlogit

model than in the Cox and logistic model alone. These additional constraints can help in prediction and feature selection. While being better at both tasks together, the Coxlogit model can also improve the Cox and logistic model in their respective tasks. This improvement is observed in particular when the predictions are difficult, that is when the number of training samples is limited or with a drastic feature selection. In those cases, the advantage of having more supervisions overcomes the problem of optimizing two objectives not perfectly correlated.

Even when it does not improve the predictive performances, the Coxlogit model is useful in feature selection. As an embedded method, the feature selection is multivariate and favors features that are good in classification and survival prediction. Moreover, having one unique model enforce a selection of features that are consistent between the two tasks. The Coxlogit model will not selected features that are positively correlated with the survival and negatively correlated with the classification.

As mixture of the Cox and logistic regression, the Coxlogit model produces risk scores that are interpretable both in terms of hazard functions and risk group probabilities. The Coxlogit model can also be viewed as a generalized linear model. It has a convex loss function which can be easily optimized, for example with an iteratively reweighted least square. Similarly as for the Cox and logistic regression, the Coxlogit model could even be kernelized to deal with non-linearity in the data [57, 84].

We have shown in this work that combining generalized models is a simple and powerful solution to exploit multiple surpervisions. Our approach could be extend to regression, combining a continuous response with the survival data. This continuous response can even be survival times. A similar approach combining regression and ranking constraints in a survival SVM was proposed by Van Belle *et al.* [138]. Their solution, however, does not allow an interpretation of the predictions in terms of hazards and is computationally heavy.

An other extension could be implemented using an ordered logistic regression. This extension could be useful to deal with more than two risk groups, for example a low, intermediate and high risk group.

The Coxlogit model can also easily be extend to data where each patient may not have both supervisions. Such situation does not even require any changes in the proposed methodolodgy (section 6.2). We can simply assign patients without labels to a specific group with label $y_i = 0$ and replace missing survival data by negative censoring times, $(t_i = -1, \delta_i = 0)$. These replacement labels and survival data will disappear in the computation of the gradient of the Coxlogit likelihood.

We can push further this idea while using the Coxlogit model simultaneously on two datasets (with no common patients): one with survival data and one with classification data. The Coxlogit model could then be used to find signatures of genes that are jointly predictive of both the survival and the class labels.

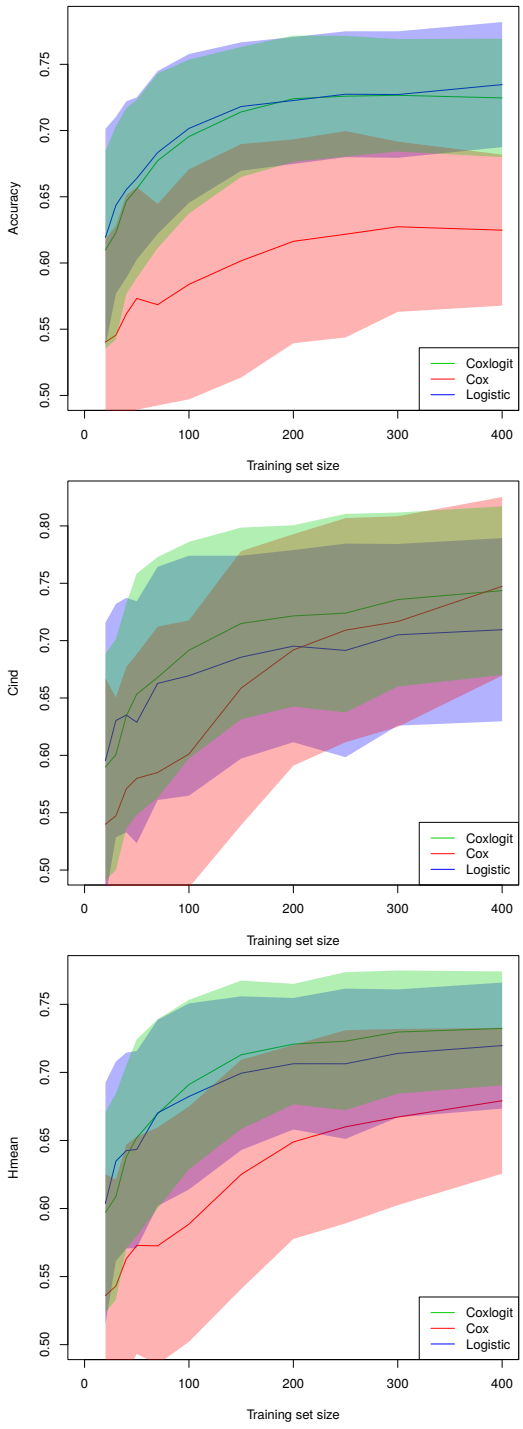


Figure 6.11: Learning curves of the Cox, logistic and Coxlogit model on breast cancer data. The feature set size is fixed to 10.

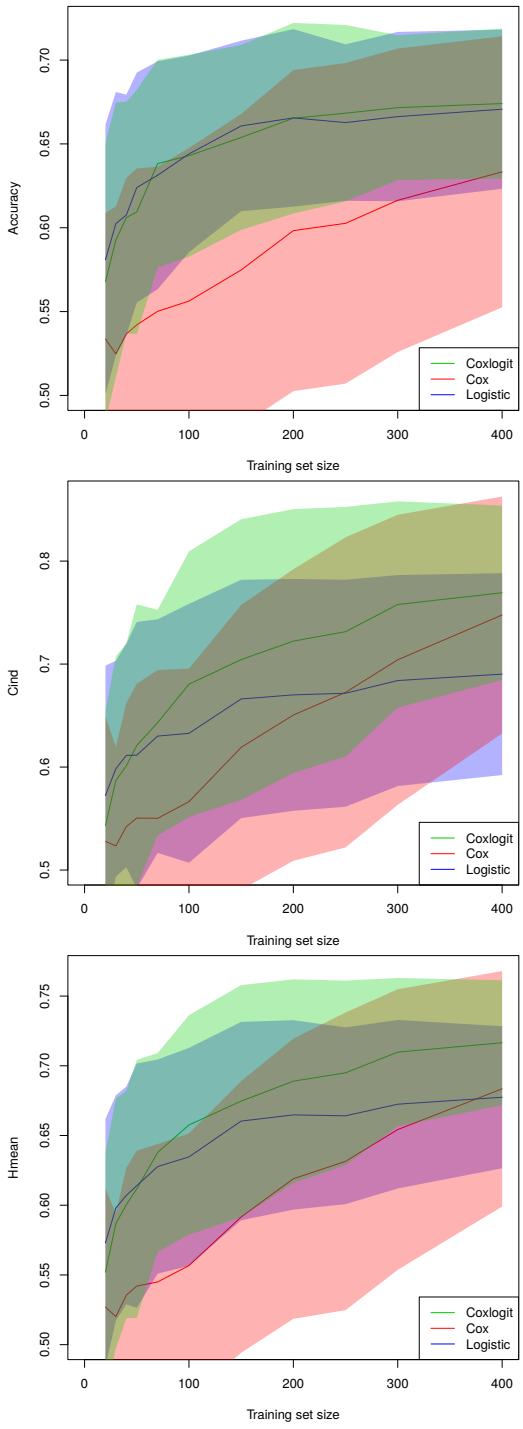


Figure 6.12: Learning curves of the Cox, logistic and Coxlogit model on breast cancer data. The feature set size is fixed to 2.

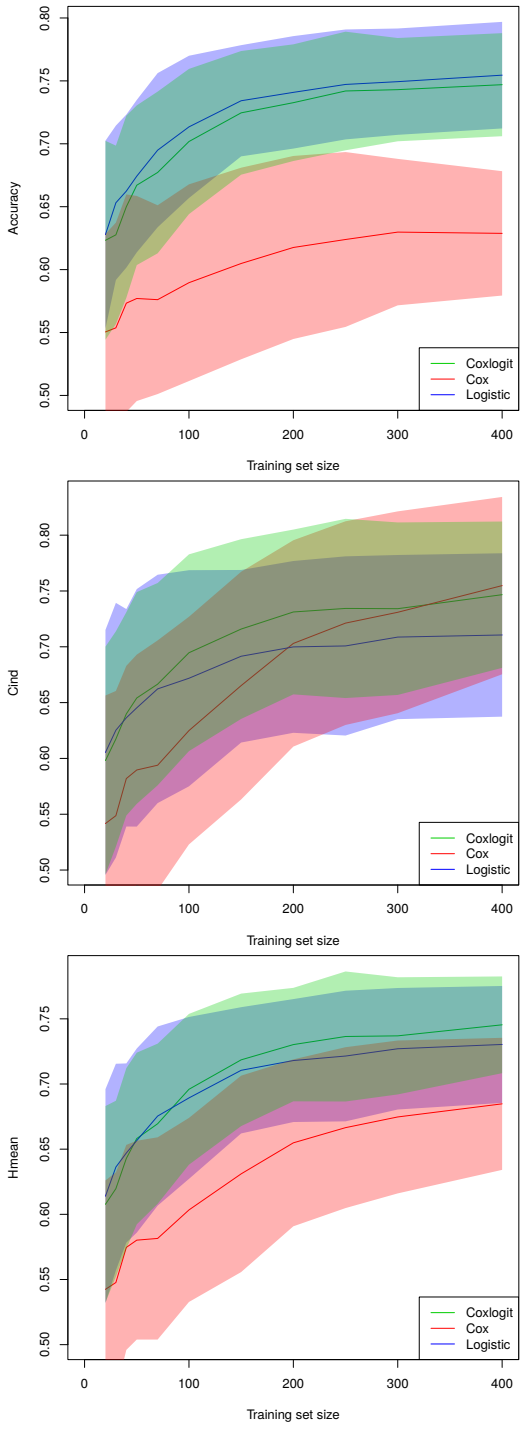


Figure 6.13: Learning curves of the Cox, logistic and Coxlogit model on breast cancer data. The feature set size is fixed to 20.

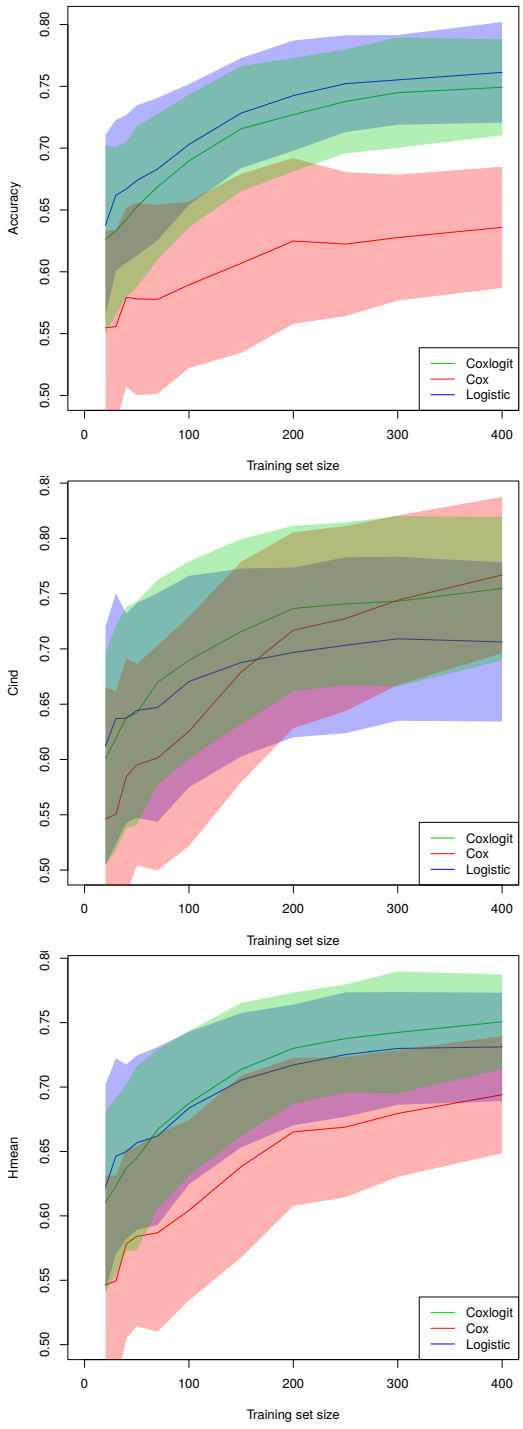


Figure 6.14: Learning curves of the Cox, logistic and Coxlogit model on breast cancer data. The feature set size is fixed to 50.

Chapter 7

Random Non-linear Projection for Survival Analysis

7.1 Introduction

In section 2.4, we show how the Cox model can be used to model and predict the survival of patients from their covariates, such as gene expression values. However, Cox models cannot handle as such non-linearities in their covariates, what may restrict their usefulness in some settings.

This chapter aims to show that Cox models can easily handle non-linear relationships if one uses random non-linear projections. Such tools have been used in extreme learning [69] to obtain results which are close to those of support vector machines, but at a much smaller computational cost. Random projections are used here with Cox models and a feasibility study is performed. Results are comparable to those of standard Cox models, but the proposed method can be used to handle data with non-linear relationships.

Section 7.2 explains what random non-linear projections are and how to use them. Section 7.3 details the proposed methodology to extend the Cox model, which is experimentally assessed in Section 7.4.

- Branders, S., Frénay, B., and Dupont, P. (2015b). Survival Analysis with Cox Regression and Random Non-linear Projections. *Proceedings of the 23th European Symposium on Artificial Neural Networks*, pages 119–124

7.2 Non-linear random projections

A potential limitation of the Cox model is that it cannot deal with non-linear relationships. Hence, a natural extension consists in adding support for features which must be non-linearly transformed to compute the hazard function. Many approaches exist in machine learning to obtain non-linear models like kernels or neural networks. However, kernelized Cox models [84] and survival-SVMs [138] come with the additional complexity of defining an appropriate non-linear kernel, whereas survival neural networks [11] are slow to learn. The proposed method focuses on a different approach which allows one to keep the interpretability and simplicity of Cox regression.

In extreme learning, it has been shown that random non-linear projections of the inputs [69] can be used to achieve state-of-the-art results in both non-linear classification and regression [68]. Those non-linear projections are obtained independently from training data: only their dimensionality p and the number of non-linear projections m must be known. The l-th projection is defined as:

$$z_l(\mathbf{x}_i) = \sigma\left(\sum_{j=1}^{p} w_{jl} x_{ij} + b_l\right) \tag{7.1}$$

where σ is a non-linear function, w_{jl} is the weight between the j-th input x_{ij} and the l-th projection and b_l is the bias used for the l-th projection. Non-linear projections could be optimized but Huang et al. [69] have shown that one can simply (i) draw the weights and biases in Equation (7.1) randomly (*e.g.* from a uniform or Gaussian distribution) and (ii) keep them fixed during learning.

The advantage of the above strategy is that state-of-the-art results are obtained in non-linear classification and regression [68] at the cost of linear methods. Indeed, the matrix of inputs $\mathbf{X}$ is replaced by the matrix $\mathbf{Z}$ of random non-linear projections, which is fixed for a given dataset and needs not be trained.

$$\mathbf{Z} = \begin{pmatrix} z_1(\mathbf{x}_1) & \cdots & z_m(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ z_1(\mathbf{x}_n) & \cdots & z_m(\mathbf{x}_n). \end{pmatrix} \tag{7.2}$$

Afterwards, fast, linear methods (linear regression, logistic model, *etc*) can be used with $\mathbf{Z}$ instead of $\mathbf{X}$. Using random non-linear projections offers a good compromise between computation needs and prediction accuracy. This view has been popularized in [86, 93, 50] where it is

shown that the number of random projections can be set to a large number (*e.g.* $m = 10^3$) or even be infinite [51] if regularization is used to control the model complexity.

7.3 Proposed methodology

We propose to use random non-linear projections $\mathbf{Z} \in \mathbb{R}^{n \times m}$ as input to a Cox model rather than the original covariates $\mathbf{X} \in \mathbb{R}^{n \times p}$. The main advantage is that non-linear relationships can now be modeled, while the interpretability of the Cox model output is preserved in terms of hazard. Also, contrarily to *e.g.* SVMs, we do not need to choose a kernel nor to tune its parameters. Since works like [86, 93, 50, 51] show that regularized linear methods work well with large numbers of random non-linear projections, L_2 regularization can be used to control the complexity of the resulting non-linear Cox model. The algorithm 7.1 is used here to generated the projections. Weights and biases are drawn from a uniform distribution between -2 and 2, and the inputs are normalized before being non-linearly transformed with an hyperbolic-tangent. Section 7.4 assesses this methodology.

Algorithm 7.1: Random non-linear projections

Data: The normalized data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, the number of random non-linear projections m.
Result: The matrix of random non-linear projections $\mathbf{Z} \in \mathbb{R}^{n \times m}$.
Generate a matrix $\mathbf{W} \in \mathbb{R}^{p \times m}$ such that $w_{i,j} \sim U([-2, 2])$;
Generate a vector $\mathbf{b} \in \mathbb{R}^{1 \times m}$ such that $b_j \sim U([-2, 2])$;
$\mathbf{Z} = \tanh\left(\mathbf{XW} + \mathbf{1}^{n \times 1}\mathbf{b}\right)$;

7.4 Experiments

This section validates the use of random non-linear projections with a Cox model. Experiments are performed on synthetic and real datasets; performances in survival regression are assessed according to the concordance index (C-index) on the risk scores (see section 4.12). The L_2 regularization constant λ of the Cox model is tuned with an internal 10-fold cross-validation on the training set.

A 10-fold cross-validation is used in all experiments with real datasets. All results are reported in forest plots containing: the average test performance in C-index for each model and the p-values of a paired t-test against the standard Cox proportional hazards model. The black squares

are centered on the average C-index. The horizontal grey lines correspond to the 95% confidence intervals.

7.4.1 Results on artificial non-linear datasets

Artificial data are considered first to assess to which extent our approach is able to deal with non-linear features. A data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is drawn from a standard distribution $\mathcal{N}(0, 1)$. The survival data (t_i, δ_i) are generated from two Weibull distributions (for event and censoring times, respectively) such that the hazard $h_i(t)$ depends on a combination $f(\mathbf{x}_i)$ of the features: $h_i(t) \propto \exp(f(\mathbf{x}_i))$. The Weibull shape parameters are set to 2.5 and 1 respectively for censoring and event times. The scaling parameters are 2914 and $20000 \exp(-1.5 f(\mathbf{x}_i)/\sigma)$, where σ is the standard deviation of $f(\mathbf{x}_i)$ over all generated samples[1]. Two non-linear combinations and one linear combination of features are considered here:

$$f_1(\mathbf{x}_i) = \sum_{j=1}^{d} x_{ij}^2 \tag{7.3}$$

$$f_2(\mathbf{x}_i) = \sum_{j=1}^{d} \exp(-x_{ij}^2) \tag{7.4}$$

$$f_3(\mathbf{x}_i) = \sum_{j=1}^{d} a_j x_{ij}. \tag{7.5}$$

Results are averaged in Figure 7.1 over 10 independent runs with $n = 1000$ instances (200 for training, 800 for validation) and $p = 5$ features. The Cox proportional hazard model is trained (i) on the 5 original features and (ii) using between 100 and 5000 random non-linear transformations of those features.

As expected, a standard (linear) Cox model is not able to deal with non-linear features (f_1 and f_2). The Cox model offers significantly better results when the original features are first transformed through non-linear random projections. Such a strategy is even not detrimental in the linear case (f_3). If a sufficiently large number (here 2000) of random projections is considered, the results are not significantly different from those of a standard Cox model. In general, the number of random projections to consider needs not be carefully tuned provided it is chosen large enough.

[1]Those values were chosen to produce events and censoring times similar to real data.

7.4.2 Results on real-world datasets

This section shows results for three real-world cancer datasets. The first real dataset is the flchain[2] dataset, which contains 8 features for 7874 patients. Multiple causes of death were recorded and the death due to a circulatory system diseases is considered here. Others causes of death are seen as censoring, which is one way to deal with competing risks [101]. The second dataset consists of five pooled breast cancer datasets from the GEO database (accession numbers: GSE2034, GSE5327, GSE7390, GSE2990, GSE11121 and GSE6532). 75% of the features with the lowest variances are removed, which is a standard pre-filtering of such high-dimensional data. The final dataset contains 1054 patients with 5571 features. The third dataset consists of seven pooled colon cancer datasets from the GEO database (accession numbers: GSE39582, GSE17536, GSE17537, GSE14333, GSE29621, GSE29623 and GSE38832). After a similar pre-filtering, the final dataset contains 1234 patients with 13669 features.

Figure 7.2 shows results obtained for the flchain, breast and colon datasets. The C-index reaches a plateau when the number of projections increases. Globally these results do not exhibit statistically significant differences with those of a standard Cox model. They illustrate that the proposed approach is effective even though explicit non-linearities might not be required for these datasets.

The sensitivity of results to the choice of the L_2 regularization parameter λ is studied in Figure 7.3 using the flchain dataset. The number of random projections is fixed to 200 and 500 and results are reported with λ equal to $\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1, 10, 20\}$ times the number of dimensions. Results are improving while increasing λ and reach a plateau, here when λ is roughly equal to the number of projections. The choice of λ seems robust and it does not seem difficult to tune.

7.5 Conclusion

This chapter shows how random non-linear projections used in extreme learning can also be used to extend Cox models. Using the proposed methodology, survival analysis can be performed even with non-linear relationships between covariates and the associated risk scores. The computational cost is comparable to the cost of learning standard Cox models. Since Cox models are essentially used to compute risk scores, the results are still readily interpretable in terms of hazards. Such an

[2]available at http://cran.r-project.org/web/packages/survival/ in the survival R package.

approach avoids the additional complexity of defining an appropriate non-linear kernel or of training complex neural networks.

One of the drawback of this approach is the model interpretation. The final model is a multivariate combination of non-linear random projections which is difficult to interpret. However, one can look at the random projections that are important in the model, *e.g.* with a high absolute weight in the model. The interesting features could be the features with a high weight in those important random projections. Preliminary results show that such an approach could be used as a non-linear feature selection.

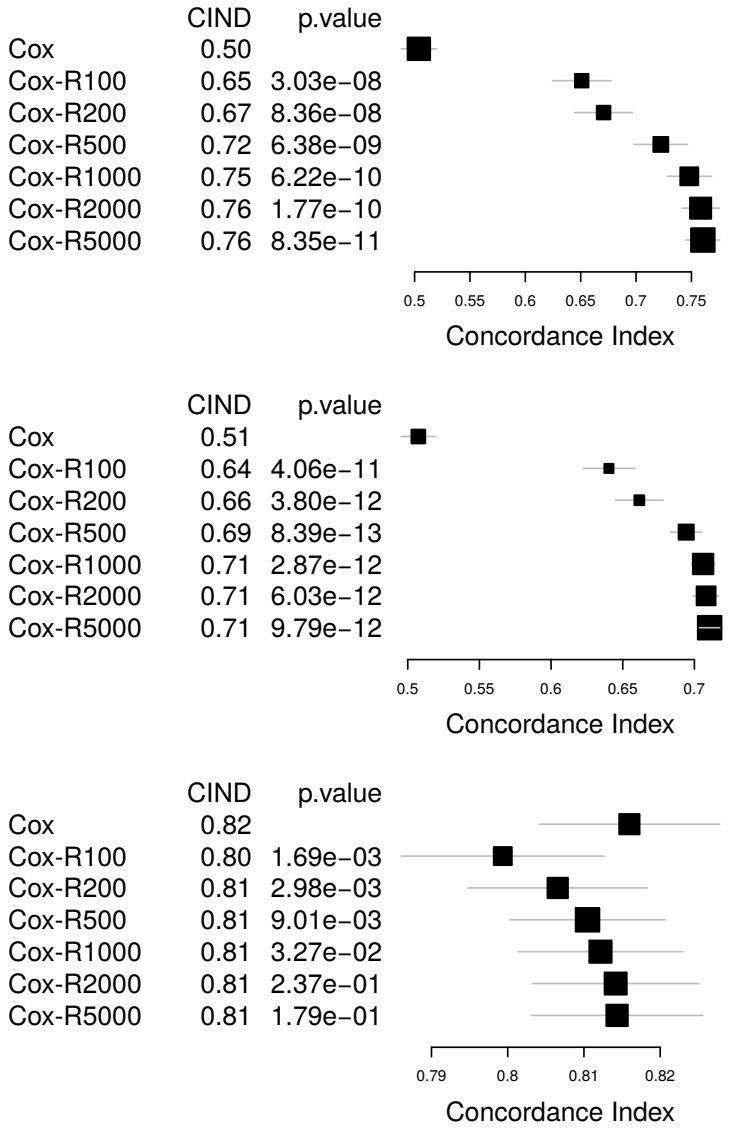


Figure 7.1: Results in C-index on synthetic data sets with $p = 5$ features. Top, center and bottom plots respectively for f_1, f_2 and f_3. The results are reported for a Cox model with either the original features (Cox) or random non-linear projections (Cox-R). The number of projections varies from 100 to 5000.

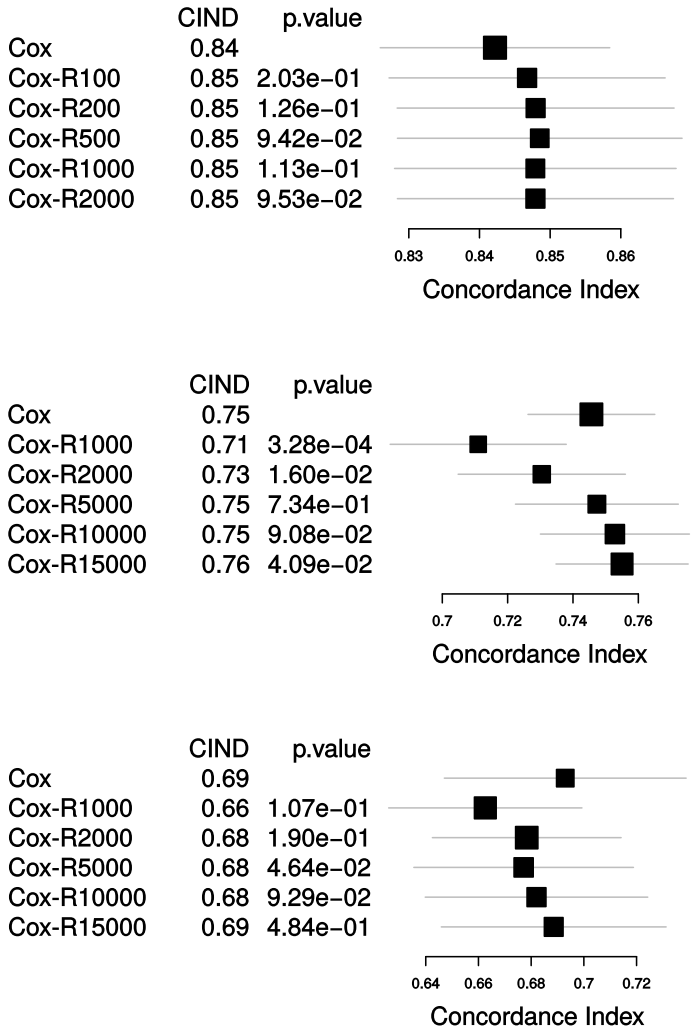


Figure 7.2: Results in C-index with the flchain, breast and colon datasets, respectively at the top, center and bottom. These three datasets have respectively 8, 5571 and 13669 original features. The results are reported for a Cox model with either the original features (Cox) or random non-linear projections (Cox-R). The number of projections varies from 100 to 15000.

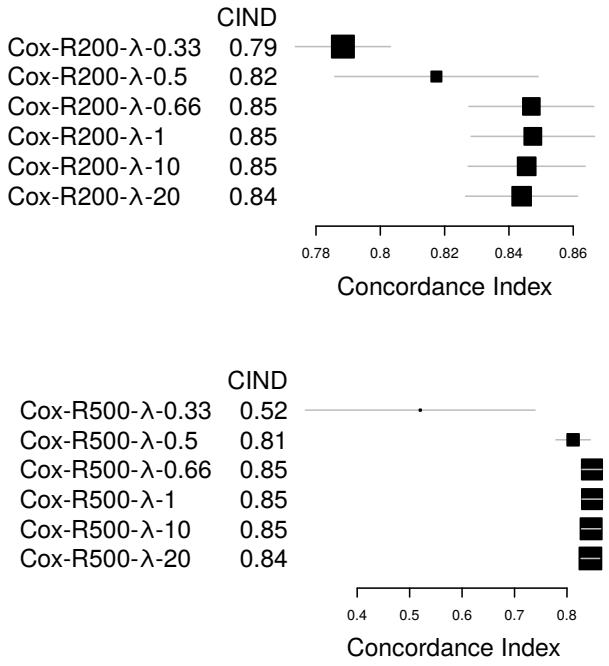


Figure 7.3: Results with the flchain dataset changing the regularization constant λ. Top and bottom plots respectively for 200 and 500 random projections.

Chapter 8

Balanced Hazard Ratio

8.1 Introduction

In chapter 4, we present several metrics used to assess risk groups and in particular how they fit with the observed survival data. In this chapter, we extend the discussion, started in section 4.2, on the limitations of these metrics. Our first contribution is to argue why the hazard ratio and C-index, while being perfectly sound to assess risk *scores*, are much less appropriate to evaluate risk *groups*.

Part of the issue comes from the choice of the threshold (or cut-off) values on risk scores to obtain the risk groups. We show in particular that the HR may be artificially increased by considering highly unbalanced groups: an extremely unbalanced choice would, for instance, consider a single patient with the shortest survival time (or the highest risk score) as the unique member of the high risk group, while putting all other samples in a presumably low risk group. Such an extreme choice is likely to lead to a very high HR but is unlikely to be valuable from an accurate prognosis viewpoint. A perfect balance, say 50%/50% between high and low risk groups, needs not be relevant either. Unless some prior information exists about the relative size of the risk groups, which is rarely the case when assessing the prognostic values of new candidate markers, the definition of risk groups looks ill-defined when assessed through the standard HR. We stress that this problem occurs beyond extremely unbalanced cases as a consequence of the HR measure exhibiting potentially many local optima and being particularly non-smooth. Therefore, very small modifications of the proportions between risk groups (through marginal cut-off modifications) can lead to highly different HR values. We also show that maximizing HR versus

minimizing its associated p-value could lead to drastically different risk groups.

Our second contribution is the definition of a new performance metric, called the *balanced hazard ratio* (BHR) to fix the issues raised above. The BHR keeps an intuitive interpretation and is as simple to compute as the original HR, meaning that it can be easily used by clinicians accustomed to the hazard ratio. Yet, the BHR penalizes artificially unbalanced risk groups and, more generally, offers a smoother profile with a natural optimum. Interestingly, such optimum is data-dependent and needs not correspond to a perfect balance between groups. Our third contribution is to show that the BHR leads to a natural definition of cut-off values on risk scores to define associated risk groups.

We illustrate the proposed methodology on breast cancer studies assessing the quality of prognostic gene signatures. Section 8.2 briefly describes those studies, which are chosen here as running examples. Yet we believe that our conclusions fully apply to the general evaluation of risk groups from survival data. Section 8.3 illustrates, on some running examples, the influence of cut-offs on the definition and the performances of risk groups. In particular, we discuss the influence of cut-offs on the original hazard ratio and why its use is problematic to assess risk groups. This discussion is extend in section 8.4 to the other performance metrics presented in section 4.2. We compare them to the hazard ratio and we discuss their relevance for evaluating risk groups. Section 8.5 presents the balanced hazard ratio. We argue why it is well designed for evaluating risk groups, while keeping a natural interpretation along the same lines as the original hazard ratio. Section 8.6 describes how the balanced hazard ratio, originally defined for two risk groups, can be easily generalized to an arbitrary number of risk groups. Section 8.7 describes how the BHR can be used to choose natural cut-off values on risk scores. Section 8.8 further illustrates the soundness of the proposed methodology on controlled experiments for which an underlying threshold between risk groups is fixed by design. We conclude our work and discuss additional perspectives in section 8.9.

- Branders, S. and Dupont, P. (2015). A balanced hazard ratio for risk group evaluation from survival data. *Statistics in Medicine*, **34**(17)

8.2 Illustrative clinical studies

Risk group prediction and evaluation are illustrated in here on known clinical studies of breast cancer prognosis. These studies offer a variety

of data sets publicly available from the GEO database with existing prognostic markers, often made of gene expression signatures. The high prevalence of breast cancer and the availability of several prognostic indexes for essentially the same task drove our choice on those examples but our conclusions are aimed to be applicable to any evaluation of risk groups from prognostic indexes and survival data. In particular, similar results with additional prognosis models for breast cancer, and further results on colon and ovarian cancers are presented in our paper [17].

The Gene76 prognostic model is built from 76 genes to identify patients who developed distant metastasis within 5 years. All patients considered are node negative and untreated. The gene expression and survival data for this study form the Veridex (VDX) data set [143]: $n = 344$ samples, GEO accession numbers GSE2034 and GSE5327.

The Gene76 prognostic model has been further validated on an independent study conducted by the TRANSBIG (TBG) consortium. The TBG data set includes untreated patients with primary breast cancer and a node negative status [35]: $n = 198$ samples, GEO accession number GSE7390.

The data set UNT comes from a study investigating the links between histopathological grades and gene expressions [124]. To focus on comparable data sets, we consider only untreated patients from this study with a node negative status after removing samples also present in VDX or TBG: $n = 84$, GEO accession number GSE2990.

In all those studies, distant metastasis is used as end point and no information is available about possible competing risks. Gene expression data are measured on the Affymetrix HGU133a microarray platform. All data sets were summarized according to the MAS5.0 procedure and represented in $\log_2$ scale. Practical experiments were conducted using the R statistical language, including specific breast cancer prognostic models implemented in the `genefu` R package [61] from Bioconductor.

8.3 Risk groups and hazard ratio

In chapter 3, we present several methods/models to compute and to predict the risk scores and risk groups of patients. These models usually define the continuous risk score of a patient i as a linear combination of its covariates $\mathbf{x}_i$:

$$r_i = \boldsymbol{\beta}^\top \mathbf{x}_i \tag{8.1}$$

The risk groups are then defined from a cut-off θ on these risk scores.

$$g_i = \text{sign}(\boldsymbol{\beta}^\top \mathbf{x}_i - \theta) \tag{8.2}$$

$$g_i = \begin{cases} -1 & \text{if} \quad r_i < \theta \\ 1 & \text{if} \quad r_i \geq \theta \end{cases} \tag{8.3}$$

Without access to observed risk groups y_i, the choice of θ can be viewed as an unsupervised or **partially supervised** problem. The survival models presented in section 3.3 compute a risk score but do not give any estimate of θ. The cut-off should be estimated in a second step to predict risk groups from the Cox or survival SVM continuous risk scores.

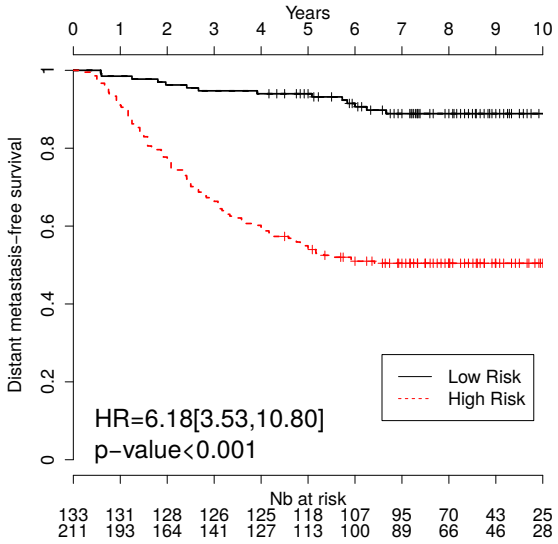


Figure 8.1: Risk groups on the VDX dataset with the original Gene76 cut-off.

To illustrate the influence of the cut-off choice, we consider the Gene76 prognostic index on the VDX dataset. This index first partitions the patients depending on their estrogen receptor status, being either positive or negative. For each status, the risk score is defined through a specific combination of univariate Cox models. Figure 8.1 reports the Kaplan-Meier curves for the 2 risk groups as defined by this model and its original cut-off. The specific details on how this original cut-off was chosen are further discussed in section 8.7. We focus here on the impact of this choice versus possible alternatives. Kaplan-Meier curves present the proportion of patients still at risk (*i.e.* without having experienced the event) along the follow-up time (expressed here in

years). Crosses on the curves represent censored data. There is one curve for each risk group with the corresponding number of patients in each group being reported below the x-axis. Informally speaking, the more separated the 2 curves, the better the prognostic index and its underlying predictors as prognostic markers. To evaluate such a difference, the hazard ratio (HR) is commonly considered, together with its 95% confidence interval and a p-value of a statistical test assessing whether HR significantly differs from 1.

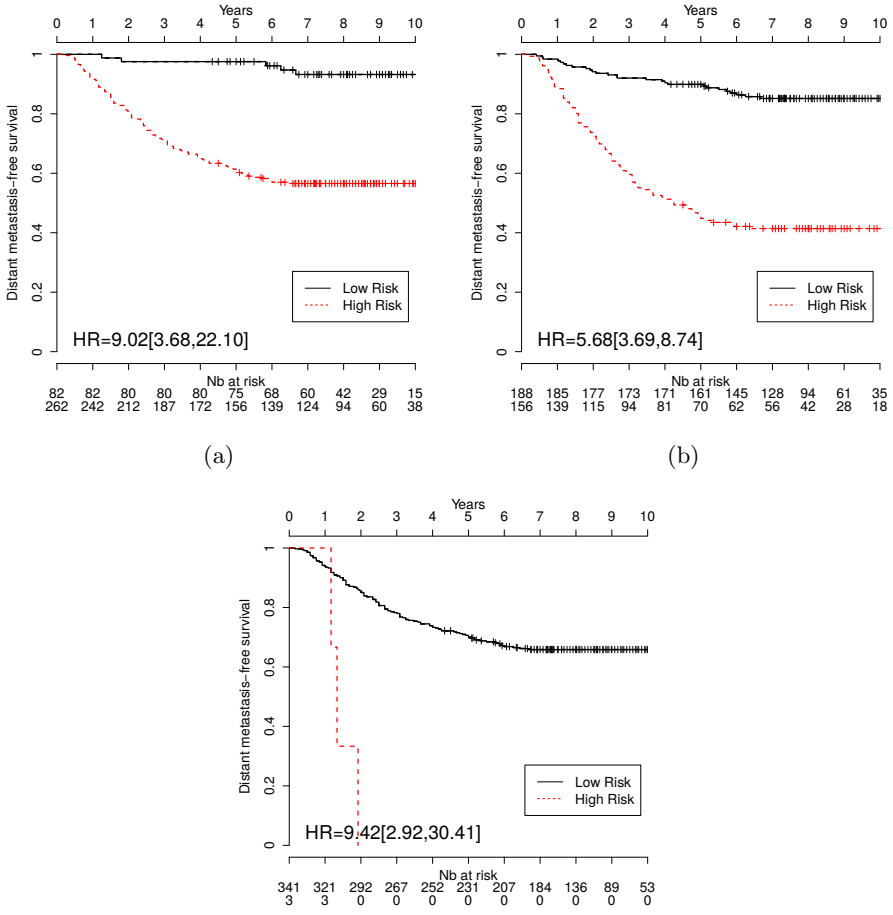


Figure 8.2: Risk groups on the VDX data set defined by the Gene76 model with alternative cut-offs.

In this regard, Gene76 looks to be a good prognostic index, as evaluated on the VDX dataset, since its HR is high (6.18), with a 95% confidence interval $= [3.53, 10.80]$ and a very small p-value. Yet, this

would need to be confirmed on additional samples independently from those used to estimate this model, as further discussed in section 8.7. Indeed, the assessment of a cut-off on the same data used for its estimation can be subject to a large bias [2].

We focus here on a different issue. Alternative cut-off choices are possible and could lead to strongly different HR evaluation. The figure 8.2 presents the survival curves of the high and low risk groups from the same Gene76 model evaluated on the same dataset (VDX) but with alternative cut-offs.

Changing the cut-offs logically affects the number of patients in the low versus high risk groups but it also largely modifies the HR values. In particular, a HR value larger than 9 is reported in figures 8.2(a) and 8.2(c), which is much higher than the original HR. In the first case, the number of patients in the low risk group has been decreased (from 133 to 82 at time 0). In the latter case, nearly all patients (341 out of 344 at time 0) now belong to the low risk group, which forms a particularly imbalanced splitting between risk groups. Such observations tend to show that an appropriate cut-off choice is critical. Before revisiting this question, we argue why assessing risk group prediction through a group hazard ratio is problematic. The results presented in figure 8.2 will be used as running examples throughout this chapter.

As illustrated in figure 8.2, a specific cut-off choice on the risk scores may largely influence the proportion of samples in each risk groups and the resulting HR values. Figure 8.3 generalizes this analysis by reporting the HR values (along the left y-axis) obtained for all possible cut-offs defining a proportion of samples in the low risk group varying from 0% to 100% (along the x-axis). Results presented in figure 8.2 correspond to 3 specific cut-offs, hence 3 specific points on this HR curve, illustrated by black, red and green dots respectively while the original Gene76 cut-off is represented with a dashed line. The HR value can clearly be artificially increased by considering extremely unbalanced risk groups, which would nevertheless be uninformative from a prognostic point of view. The problem is even more serious since this HR (plain) curve exhibits many local optima and is far from being smooth. In other words, marginal changes in the relative proportions between risk groups may drastically affect the observed hazard ratio while potentially modifying, positively or negatively, the estimated quality of the predictors used as prognostic markers. Finally, due to those fluctuations, the evolution of the HR values are inconsistent with the associated p-values (dashed orange line, $-\log$ scale along the right y-axis). Hence risk groups maximizing the HR value (*e.g.* the green dot) is largely different from those minimizing the associated p-values (`e.g.` the red dot).

All the above results illustrate that using the hazard ratio to assess the quality of prognostic markers to discriminate between risk groups is highly questionable.

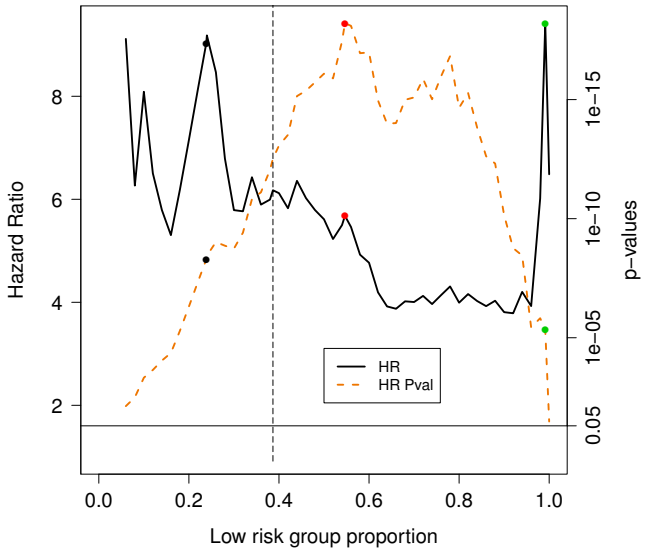


Figure 8.3: Plain line: hazard ratio of the Gene76 model on the VDX dataset while varying the proportions in each risk group through different cut-off choices (HR value on the left y-axis). The dashed vertical line corresponds to the original cut-off. Dashed orange line: associated p-values plotted in $-\log$ scale (right y-axis, the higher the better on such a plot).

We stress that the issues raised here directly follow from the use of a **discrete** indicator variable g_i in the definition of the group hazards (section 4.2.1). Another classical definition of HR considers a **continuous** risk score r_i instead of the discrete (and here binary) g_i. In the continuous case, let us assume for example $HR = 2$ while comparing the risk scores $r_i > r_j$, associated to patients i and j still at risk. Such a HR value would mean that the probability of experiencing the event is twice as large for patient i. This is a perfectly sound use of the HR measure to assess the relevance of risk scores. The problems raised above appear whenever cut-offs are chosen on those scores and discrete groups are defined accordingly. Such discretization needs however to be considered as it is routinely used by clinicians to decide whether a specific treatment should be given to a patient. Such decision is indeed often based on

the assignment of the patient to a particular risk group. Royston *et al.* [110] argue that this discretization should only be applied on risk scores at the very end of the model-building process. The dichotomization of the continuous predictor variables may indeed introduce problems such as loss of information, reduction in power, uncertainty in defining the cutpoints.

We introduce in section 8.5 a novel performance metric, called the balanced hazard ratio, which fixes those issues while keeping an interpretation similar to the original hazard ratio. In the meantime, we extend the previous results to alternative performance metrics.

8.4 Alternative performance metrics

In section 4.2, we presented several performance metrics proposed in the literature to evaluate risk group prediction models, including the concordance index, the logrank test, the SEP and the pair sensitivity and specificity.

We study here the relative behaviors of these various performance metrics on our running example. Figure 8.4 further extends figure 8.3 by reporting the five metrics under study after rescaling all of them between 0 and 1 to ease the comparison (BCR is reported at 5 years after treatment, for all possible risk score cut-offs).

We note that the C-index behaves very similarly to the hazard ratio: one can trivially optimize them while considering artificially unbalanced groups, they both exhibit many local optima and sharp fluctuations for marginally different group proportions. In contrast, the logrank test, the SEP and the BCR look more appropriate as they offer quite smoother curves with a similar global optimum observed for more balanced groups. We note that this optimum is data dependent and needs not correspond exactly to a 50%/50% balance between groups. Yet those measures are not fully satisfactory either. The logrank statistics is a sum of hypergeometric random variables which does not offer an easy interpretation in terms of survival times differences between risk groups, unlike the hazard ratio and the C-index. Yet, actual survival times are key features for the patients and for choosing appropriate treatments.

The SEP metric offers a more direct interpretation than the logrank and provides an estimate of the degree of separation of the different risk groups. SEP and logrank behave very similarly but also share a common disadvantage: they are insensitive to the ordering of the risk groups. In other words, those values are unchanged after inverting risk groups and wrongly assigning the patients with a lower risk score to the higher risk

group. This problem becomes even more serious with more than 2 risk groups (see section 8.6).

As for the BCR, it is not a distance measure between survival curves but rather a balanced measure of classification rates in each group. It also depends on a critical time value (here chosen at 5 years after treatment) which is somewhat arbitrary and moreover largely dependent on the pathology under study. Finally, specificity/sensitivity measures are not perfectly suited to censored data because they simply ignore the patients who did not experience the event and have been censored before the critical time considered.

Section 8.5 introduces a novel performance metric to address those issues.

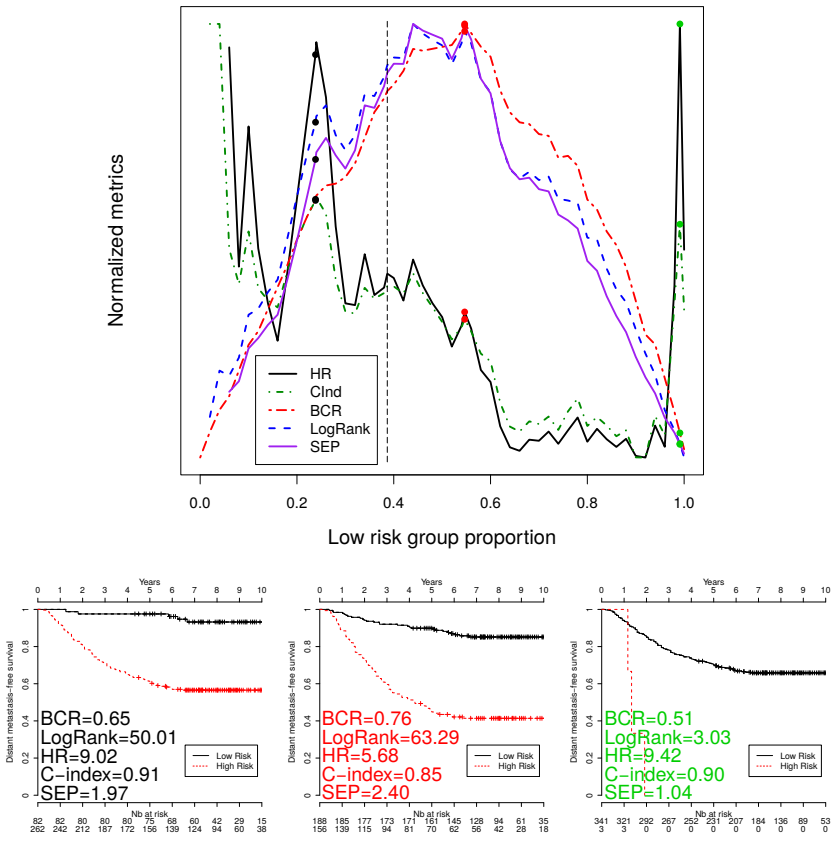


Figure 8.4: The various performance metrics evolution while varying the proportions in each risk group through adjusting the cut-off on risk scores. The original Gene76 cut-off corresponds to the dashed line.

8.5 Balanced Hazard Ratio

The balanced hazard ratio (BHR) computes the hazard ratio between three curves: the survival curves of the high and low risk groups (as for HR) and a third global survival curve over all patients (we present a generalization to more than 2 risk groups in section 8.6). Each sample is now considered as a member of 2 groups: its actual risk group ($g_i = -1$, for low risk, or $g_i = 1$, for high risk) and the global risk group ($g_i = 0$) for all patients. Such a global risk group represents the hazard (or survival time) over the whole population of patients and one measures now how much each specific risk group departs from the global curve. Figure 8.5 illustrates those survival curves on our running example with the 3 different proportions between risk groups considered so far.

The hazard function is now defined over those 3 groups:

$$h_i(t) = h_0(t) \exp(\beta g_i), \text{with } g_i = -1, 0, \text{ or } 1 \tag{8.4}$$

The quantity $h_0(t)$ represents here the hazard of the whole population, $h_0(t)/\exp(\beta)$ the hazard of the low risk group and $h_0(t)\exp(\beta)$ the hazard of the high risk group. The balanced hazard ratio, $BHR = \exp(\beta)$ is simply the multiplicative factor to get the hazard of the high risk from the global hazard or from the low risk to the global one. We also note that the square of this value, $[\exp(\beta)]^2$, has the same scaling as the original hazard ratio and can be interpreted as the relative hazard between the high and low risk groups. However, note that if BHR^2 has a similar interpretation, it is not equal to the HR. Indeed, the β value is found by fitting a partial likelihood according to the 3 curves (see below).

According to the BHR formulation, whenever the vast majority of patients are artificially considered in one group, the difference between the global survival and the survival of this group will be small (see, for example, Fig. 8.5(a) and Fig. 8.5(c)). As such, the BHR penalized extremely unbalanced risk groups without forcing risk groups to be of equal size. The consideration of the global survival curve also has a smoothing effect on the BHR because a change in survival times for one specific group only affects the hazard ratio between this group and the global survival curve.

Figure 8.7 (a) further details our running example while reporting the BCR, logrank, HR and BHR (we left out the C-index and SEP for clarity as they behave like the HR and the logrank, respectively). The BHR exhibits a behavior similar to the BCR and the logrank while offering a natural interpretation in terms of survival differences between groups as

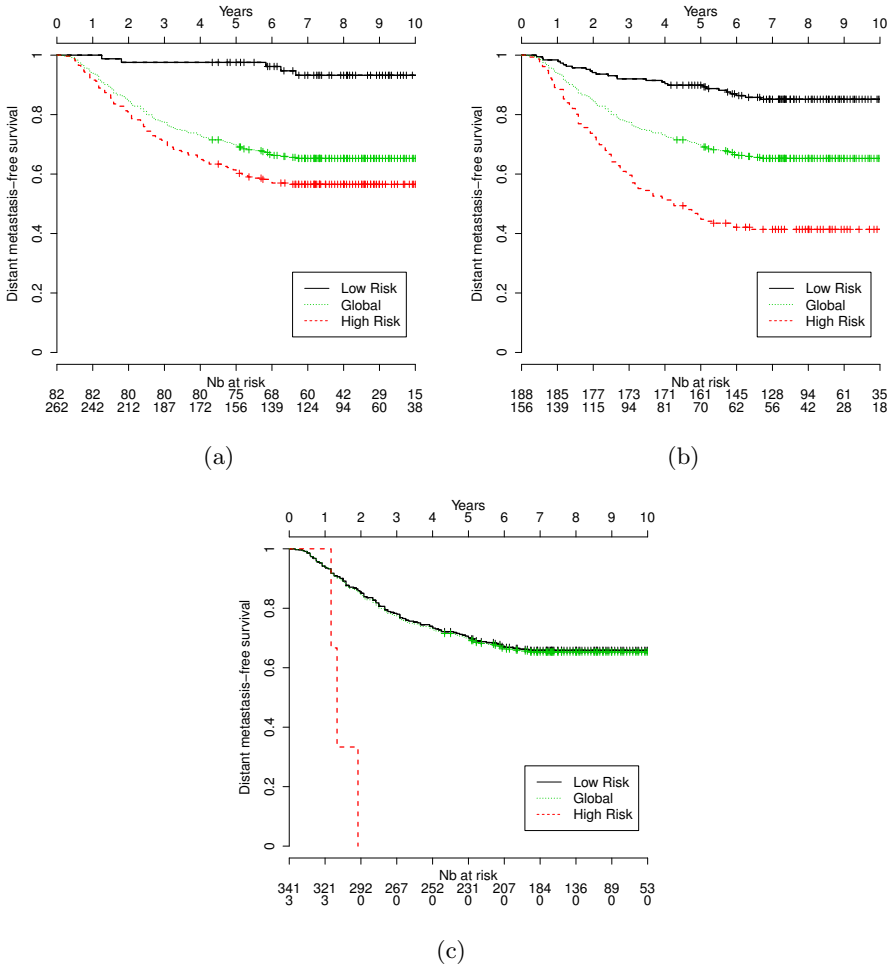


Figure 8.5: Risk groups with 3 different cutoffs and the global survival curve (in green) for the Gene76 model on the VDX dataset.

the original HR. In particular, BHR is much smoother than HR, exhibits a data-dependent global optimum and penalizes artificially unbalanced groups. Those results are extend in appendix C while comparing the hazard ratio, SEP and D-index to the balanced hazard ratio.

The smoother profile of the BHR compared to the HR can be explain by the differences in the size of their confidence intervals. Figure 8.6 presents the BHR and the HR with their 95% confidence intervals. Here, the balanced hazard ratio is squared to be on the same scale as the hazard ratio and to ease comparison. The confidence interval is much bigger for the hazard ratio than the BHR^2. Those problems of variability and extreme values of the hazard ratio occurs beyond extremely unbalanced cases, *e.g.* the HR confidence interval is already twice as big as the BHR^2 confidence interval with a low risk group proportion of 0.4.

Unlike the logrank and the SEP, the BHR is sensitive to the risk group ordering: inverting the risk groups would lead to a BHR value below 1 while BHR tends to 1 when the survival differences between the risk groups tend to vanish.

The estimation $\hat{\beta}$ of the β value from the balanced hazard ratio (see equation (8.4)) is computed through the maximization of a partial likelihood, similarly to the original HR [32]. For the BHR, the partial likelihood is slightly modified to include the global survival without duplicating the patients. The partial likelihood for the balanced hazard ratio (with the standard Breslow approximation for ties [28]) is:

$$L = \prod_{i=1}^{n} \left[\frac{\exp(\beta g_i)}{\left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1 \right)^2} \right]^{\delta_i} \text{, with } g_i = -1 \text{ or } 1 \tag{8.5}$$

The $\hat{\beta}$ value has most good properties of a maximum likelihood estimate and has an asymptomatically normal distribution. The variance of $\hat{\beta}$ can be estimated with the inverse of the Fisher information [32] and can be estimated through the second derivative of the log-likelihood with respect to β:

$$var(\hat{\beta}) \approx - \left(\frac{d^2 logL(\hat{\beta})}{d\beta^2} \right)^{-1} \tag{8.6}$$

One can thus easily compute confidence intervals and use standard statistical tests (Wald, likelihood ratio, score test [28]) to assess to which extent the BHR significantly departs from 1.

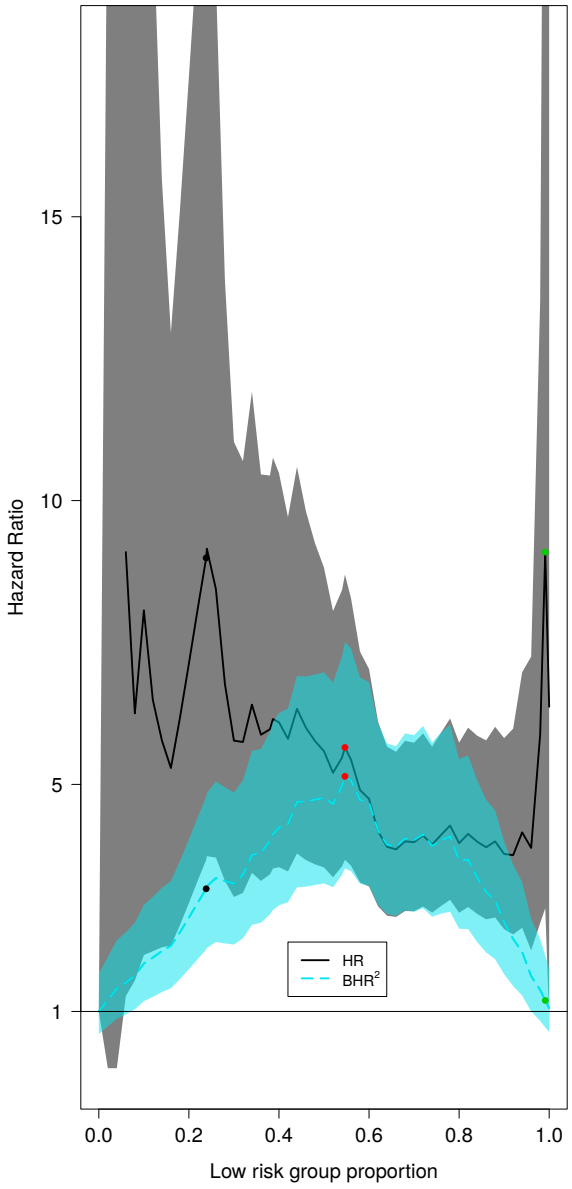
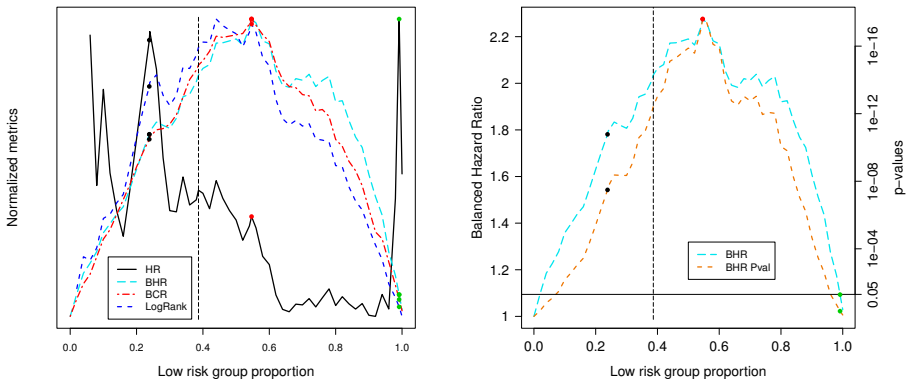


Figure 8.6: Hazard ratio, squared balanced hazard ratio and their confidence intervals on the VDX dataset while varying the proportions in each risk group. The balanced hazard ratio is squared to be on the same scale as the hazard ratio.

Figure 8.7 (b) illustrates that the associated p-values are fully concordant with the BHR values: an increase of BHR goes along a decrease of the associated p-value (here plotted in $-\log$ scale). Those results drastically contrasts with those obtained for the original HR (see figure 8.3).



(a) (b)

Figure 8.7: (a) Performance metrics on the VDX dataset while varying the proportions in each risk group. The original Gene76 cut-off corresponds to the dashed line. (b) Dashed orange line: p-values associated to BHR plotted in $-\log$ scale (right y-axis, the higher the better on such a plot)

To compare two prognostic models on the same data, we compare their BHRs with a t-test for dependant samples, similarly as in [60].

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\hat{s}_1^2 + \hat{s}_2^2 - 2r\hat{s}_1\hat{s}_2}} \tag{8.7}$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the two estimated log balanced hazard ratios, $\hat{s}_1^2$ and $\hat{s}_2^2$ are their estimated variances. r is the Spearman's rank correlation coefficient between the predicted risk groups (*i.e.* without the global risk group) of the two prognostic models. Under the null hypothesis that the estimated BHRs are equal, the statistic (8.7) follows a student distribution of $n-1$ degrees of freedom, where n is the original number of patients.

8.6 BHR generalized to more than two risk groups

The original balanced hazard ratio (see equation (8.4)) is formulated with 2 risk groups and an additional global group representing all patients. The BHR can be easily generalized to an arbitrary number of original risk groups. The k original risk groups are assumed to be ordered from lower to higher risk and arbitrarily numbered $1, 3, \dots 2k-1$. For each pair $(i,\ i+2)$ of consecutive risk groups following this order, one additional risk group numbered $i+1$ is considered gathering the patients of this pair of groups. In total, one considers $2k-1$ (original and additional) risk groups. The hazard function is now defined as follows:

$$h_i(t) = h_0(t) \exp(\beta g_i), \text{with } g_i \in [1 : 2k-1] \tag{8.8}$$

The balanced hazard ratio, $BHR = \exp(\beta)$ is the multiplicative factor between the hazards of two consecutive groups i and $i+1$. This formula is equivalent[1] to the original BHR definition whenever $k = 2$.

The SEP and BHR metrics are similar when restricted to 2 groups as they both compare the survival of each risk group to the global survival (yet only BHR is sensitive to the group ordering). When considering a larger and arbitrary number k of groups, those metrics differ more strongly since the BHR introduces multiple new groups while SEP only compares the survival in each (arbitrarily ordered) risk group to the global survival.

The BHR can also be extended to continuous risk scores. Indeed, the limit case consists in assigning one single patient to each risk group. Each risk group would then be representative of a specific risk score and those groups can be sorted accordingly. While this is a natural limit case, it does not offer specific advantages over the HR computed on those risk scores. The purpose of introducing the BHR is to address the problems of the original HR whenever continuous risk scores are discretized into risk groups, as discussed in section 8.3.

8.7 Cut-off choice and risk group prediction

A relevant cut-off value is necessary to define risk groups from continuous risk scores, as formalized in equation (8.3). The choice of a specific cut-off is not always clearly motivated in the literature and sometimes not even explicitly described. However, given its potentially critical effect

[1] up to an arbitrary shift in group numbering.

on the estimated quality of a prognostic model and its associated prognostic markers, it looks important to use an appropriate methodology to fix cut-offs. For instance, the original cut-off associated to the Gene76 prognostic index [143] has been chosen to achieve 100% of sensitivity and the highest specificity on the training set (an undisclosed fraction of the VDX dataset). Whether such choice is really optimal for the accurate prognosis on independent samples requires an additional validation as detailed below. In any case, even on the VDX dataset from which the Gene76 model has been estimated, the dashed line on figure 8.7 illustrate that such cut-off is arguably sub-optimal. Besides, choosing among all possible cut-offs the one minimizing the associated p-value underestimates such p-value due to the multiplicity of the test [2].

The BHR measure offers a natural way to choose a relevant cut-off and the associated risk groups. One typically considers a training and a validation set. Such validation set can be made of independent samples from the same study or, preferably, from an independent study on the same medical question. Such a scheme can even be generalized to a cross-validation protocol or while using several independent resamplings from various related clinical studies. In any case, the training data should typically be used to estimate the parameters of a prognostic model (*i.e.* the identity of the prognostic markers and the way to combine their values in a single risk score) as well as the cut-off on the risk scores. In particular, unless there is some prior knowledge on the relative size of each risk group, we recommend to choose the cut-off maximizing the BHR on the training set.

We illustrate the proposed methodology with the Gene76 model while changing its cut-off to maximize BHR on the VDX data from which it was originally estimated. We compare the predictive performances obtained on independent samples for various cut-off choices made on the training data. The BHR cut-off precisely corresponds on the training set (here the VDX data) to the risk group proportion defined by the red dot on figure 8.7. The original Gene76 cut-off is represented by the dashed line while the cut-off optimizing HR is represented by a green dot.

Figure 8.8 illustrates the impact of those 3 cut-off choices on independent samples for breast cancer prognosis. Their predictive performances were assessed in terms of BCR, hazard ratio, balanced hazard ratio, C-index and logrank. We consider in particular the TBG [35] and UNT [124] clinical studies as an independent validation set.

Figure 8.8(a) reports the survival curves of the risk groups resulting from Gene76 with the cut-off optimizing BHR on the training. Such choice leads to the best validation results. In contrast, choosing the cut-

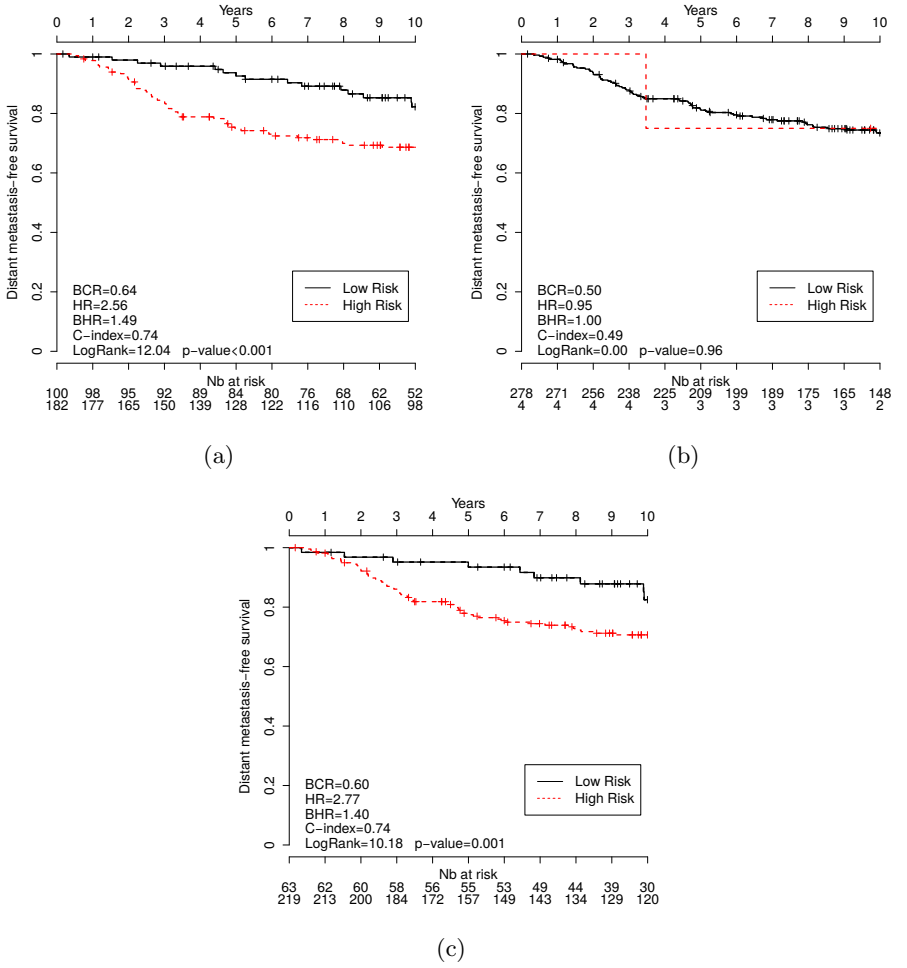
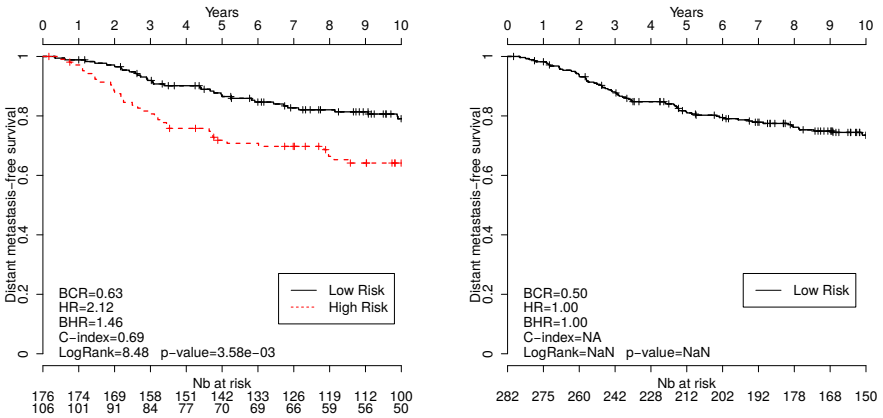


Figure 8.8: Prognostic performances on an independent validation set (TBG and UNT datasets) according to various cut-off choices on the training (VDX): largest BHR (a), largest HR (b), and the original Gene76 cut-off (c).

off to optimize the HR on the training leads to artificially unbalanced groups both on the training and on the validation set (Figure 8.8(b)) and much degraded validation performances. Figure 8.8(c) reports the validation results of the Gene76 model using its original cut-off. Those results are sub-optimal as well and illustrate that the Gene76 model could be made more effective by our proposed methodology to define risk groups. In particular the original cut-off leads to lower BHR, BCR, SEP and logrank values, which are the 4 metrics not favoring artificially unbalanced risk groups.

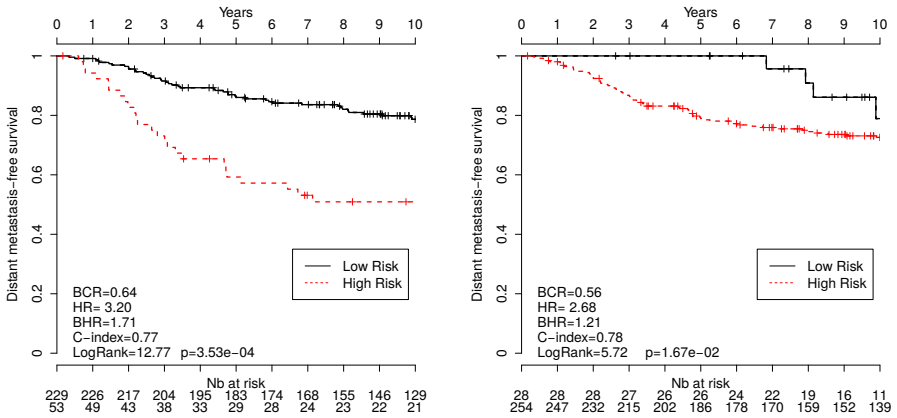
Figures 8.9 and 8.10 confirm these results on the same breast cancer data with two other prognostic models, respectively Gene70 and CoxTtest. Gene70 specifically refers to the model described in van 't Veer *et al.* [141]. CoxTtest is a multivariate Cox model built on the 100 most differentially expressed genes in the VDX dataset. For the estimation of this gene signature, two conditions are defined (event observed or not before the critical time point). We chose here 5 years after treatment as commonly accepted for breast cancer studies.



(a) BHR cutoff (b) HR cutoff

Figure 8.9: Prognostic performances on an independent validation set (TBG and UNT datasets) according to two cut-off choices on the training (VDX): largest BHR (a) and largest HR (b). Prognostic model: Gene70.

The proposed methodology to fix a cut-off maximizing BHR on the training set is further validated with other prognosis models and other cancer studies (on the supplementary materials of our paper [17]). In all cases, those results illustrate the benefits on independent validation samples of considering BHR instead of HR for fixing those cut-off values.



(a) BHR cutoff (b) HR cutoff

Figure 8.10: Prognostic performances on an independent validation set (TBG and UNT datasets) according to two cut-off choices on the training (VDX): largest BHR (a) and largest HR (b). Prognostic model: CoxTtest.

This methodology is also assessed on controlled experiments described in the next section.

8.8 Controlled experiments

The results presented in section 8.7 show that maximizing the BHR to choose a cut-off on risk scores is a good approach to optimize prognosis performances on new and independent samples. We further motivate this approach in controlled experiments for which an underlying threshold between risk groups is **a priori** fixed according to a prescribed proportion between a high or low risk profile. We assess in particular to which extent the proposed methodology is able to recover the appropriate proportions between risk groups and hence the underlying cut-off to be set on risk scores.

Synthetic data are generated for n patients separated into a low ($g_i = -1$) or high ($g_i = 1$) risk group according to prescribed proportions, ρ and $1-\rho$ respectively. Survival data and risk scores (t_i, δ_i, r_i) are randomly generated for each sample $i \in [1, n]$. The time to event t_{e_i} of patient i is drawn from a Weibull distribution and the censoring time t_{c_i} is drawn from an exponential distribution:

$$t_{e_i} \sim \text{Weibull}(\lambda_1 \exp(-\frac{\mu g_i}{2k}), k) \tag{8.9}$$

$$t_{c_i} \sim \text{Exp}(\lambda_2) \tag{8.10}$$

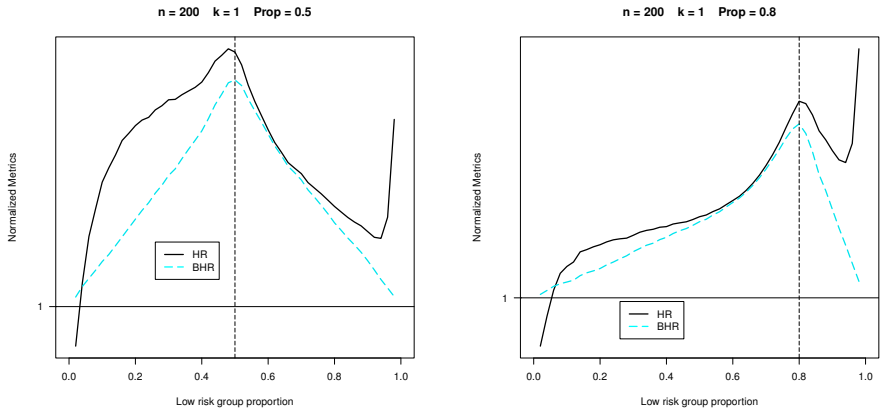
The scaling parameters $\lambda_1 = 0.002$ and $\lambda_2 = 0.01$ are fixed according to [108] from which these experiments are inspired. The shape parameter k varies in $[0.5, 1.5]$ in our experiments. Patient i is censored ($\delta_i = 0$) if his censoring time t_{c_i} occurs before the time to event t_{e_i} and the time t_i is simply defined as the minimum between both times: $t_i = \min(t_{e_i}, t_{c_i})$. The true hazard ratio between groups can be directly controlled using this protocol since it is given by $HR = \exp(\mu)$.

The risk score of a patient i is drawn from a Normal distribution centered on g_i, -1 or 1, respectively for low or high risk group. The risk scores are then distributed according to a mixture of the two Normal distributions, according to the prescribed proportion ρ between risk groups: $r_i \sim \rho\mathcal{N}(-1, 0.5) + (1 - \rho)\mathcal{N}(1, 0.5)$. The underlying threshold to be discovered is defined as the ρ-percentile of this distribution. We note that a perfect discrimination between risk groups could hardly be obtained since risk scores overlap across risk groups, as expected in a real scenario. Results are reported below over 500 independent runs of such controlled experiments.

Figure 8.11 reports the hazard ratio (HR) and balanced hazard ratio (BHR) computed while varying the cut-off used to define the risk groups. Results are reported here for two prescribed proportions $\rho = 50\%$ or $\rho = 80\%$ of the low risk profile but the same conclusions can be drawn from other ρ values. In particular, maximizing BHR leads to chose a cut-off on risk scores which, when averaged over 500 runs, corresponds to the correct underlying proportion between risk groups. In contrast, maximizing HR may lead to an inappropriate cut-off choice favoring strongly unbalanced groups.

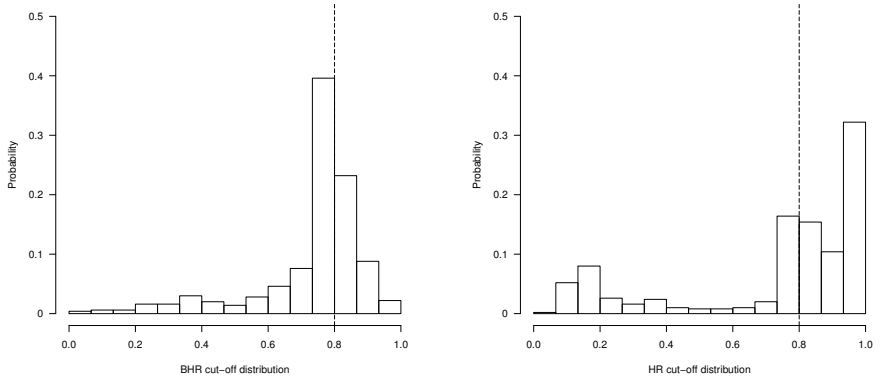
Figure 8.12 offers a closer look at the distribution over these 500 runs of the proportions between risk groups for which BHR, respectively HR, is maximum. The true proportion ρ in the low risk group was here fixed to 80%. The maximal BHR is clearly more concentrated around the true underlying proportion while the maximum HR distribution is much more dispersed and skewed towards an excessively large value. Similar results are presented using the D-index and SEP metrics in appendix C.

Figure 8.13 generalizes the above analysis while changing the shape parameter k of the Weibull distribution used to generate the survival data. It illustrates that selecting a cut-off value while maximizing HR would be even more inappropriate as k is increased to 1.5 (see, in par-



(a) (b)

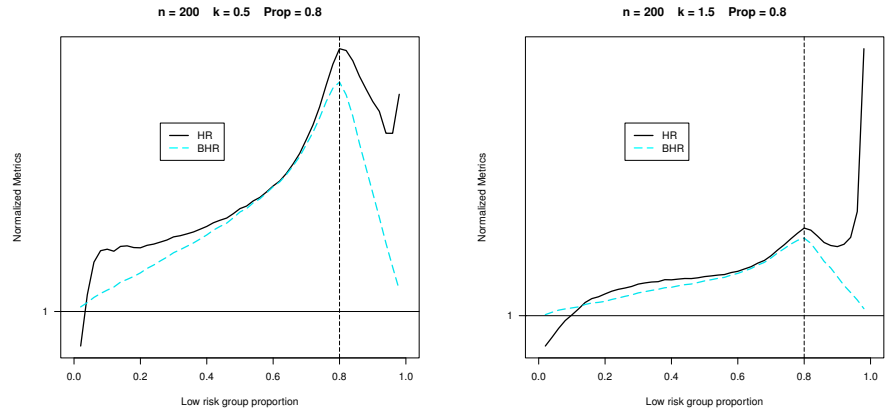
Figure 8.11: Evolution of HR and BHR, averaged over 500 runs, while varying the proportions in each risk group through adjusting the cut-off on risk scores. The experiments are conducted with $n = 200$ patients, the shape parameter $k = 1$ and the true group hazard ratio $\exp(\mu) = 3$. The true proportion ρ of patients in the low risk group was set to 50% (Figure 8.11 (a)) or 80% (Figure 8.11 (b)).



(a) (b)

Figure 8.12: Distribution over 500 runs of the low risk group proportion for which BHR (a) respectively HR (b), is maximum. The experiments are conducted with $n = 200$ patients, the shape parameter $k = 1$ and the true group hazard ratio $\exp(\mu) = 3$. The true underlying proportion of patients in the low risk group was set to 80%.

ticular, Figure 8.13 (b)). In contrast, maximizing BHR remains an appropriate methodology across various shape values.



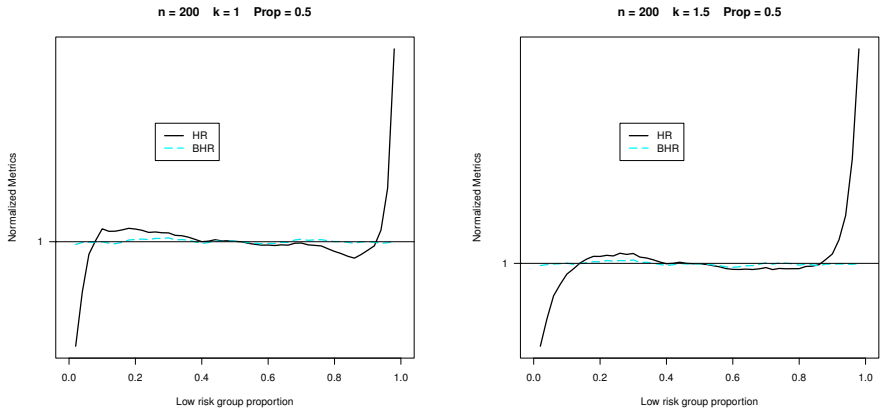
(a) (b)

Figure 8.13: Evolution of HR and BHR, averaged over 500 runs, while varying the proportions in each risk group through adjusting the cut-off on risk scores. The experiments are conducted with $n = 200$ patients, the true proportion $\rho = 0.8$ in the low risk group and the true group hazard ratio $\exp(\mu) = 3$. The shape parameter k of the Weibull distribution was set to 0.5 (a) or 1.5 (b).

Our final experiments is considering a true group hazard ratio $\exp(\mu)$ equal to 1. In other words, the data is generated such that there is actually no survival differences between both groups. Figure 8.14 reports the averaged BHR and HR values over 500 runs while changing the cut-off on risk scores. The flat BHR curve illustrates that no specific cut-off should be chosen here and hence all patients should be assigned to a common risk group. In contrast, when maximizing HR, a dichotomization into highly unbalanced groups is again promoted.

8.9 Conclusion and perspectives

Properly assessing risk group prediction from survival data is essential to analyze the relevance of candidate prognosis markers. We show here that the group hazard ratio (HR) and the concordance index, often used in such a context, can be inappropriately optimized by considering artificially unbalanced risk groups. They also exhibit many local optima and non smooth behaviors which make their evaluation highly sensitive to small fluctuations. Alternative existing metrics include the logrank



(a) (b)

Figure 8.14: Evolution of HR and BHR, averaged over 500 runs, while varying the proportions in each risk group through adjusting the cut-off on risk scores. The experiments are conducted with $n = 200$ patients, $\rho = 0.5$ and a true group hazard ratio $\exp(\mu)$ equal to 1 (no survival difference between groups). The shape parameter k of the Weibull distribution was set to 1 (a) or 1.5 (b).

statistics, the SEP metric and the average between sensitivity and specificity, also called BCR. While the logrank is harder to interpret in terms of survival differences between risk groups, the BCR is not fully adequate to censored data and relies on an additional critical timepoint that is highly dependent on the pathology. Moreover, the logrank and the SEP are insensitive to a specific ordering of the risk groups. They measure survival differences between risk groups but not the orientation of those differences.

We present here the balanced hazard ratio (BHR) which has a similar interpretation as the original HR, arguably the most common metric used by clinicians and bio-statisticians to assess risk groups. The BHR penalizes extremely unbalanced risk groups and, more generally, offers a smoother profile with a natural optimum. Its value is much less influenced by marginal changes in the proportions between risk groups and its behaves consistently with its associated p-value.

We further show how the BHR may be generalized to an arbitrary number of risk groups and how it can be used to choose an appropriate cut-off on risk scores to define risk groups. Such a methodology is both simple computationally and is shown to be sound in controlled experiments as well as real clinical studies.

Our future work includes the design of estimation algorithms for prognosis models optimizing BHR directly. Currently, Cox hazards models are typically estimated to optimize the partial likelihood on the training data. It looks preferable to fit parameters optimizing directly the final performance metric of the prognostic model. We have shown here that optimizing the group hazard ratio would be largely inappropriate. Fitting a model to optimize the BHR instead looks to be a promising alternative.

The balanced hazard ratio could also be extended to competing risks. Competing risks are modeled either through cause-specific hazards [101] or cumulative incidence functions [49]. The other causes of events are dealt with in such models either through censoring or as events occurring at an infinite time. Similarly to our proposal, a third global risk group could be introduced in these models to penalize extremely unbalanced risk groups.

An online tool and a R package are in development to make the balanced hazard ratio available to any clinician or bio-statistician.

Chapter 9

Balanced Hazard Ratio p-value

9.1 Introduction

In this chapter, we detail how p-values can be computed respectively for the hazard ratio (HR) and the balanced hazard ratio (BHR) [17] using a score test statistic. Both HR and BHR are computed here according to two risk groups defined as $g_i = \{-1, 1\}$. This definition of the risk groups gives a consistent scaling between both measures with no loss of generality. In such a setting we show that the BHR p-value is conservative with respect to the HR p-value.

9.2 Hazard ratio

The group hazard ratio (HR) evaluates the difference between survival curves computed by a Cox proportional hazards model. It represents the increase in the risk of event between the low and high risk groups. When used to evaluate risk groups, the hazard ratio is computed with a Cox model using the binary group variable $g_i \in \{-1, 1\}$ as single covariate. The hazard function $h_i(t)$ for a patient i is then written as:

$$h_i(t) = h_0(t) \exp(\beta g_i) \tag{9.1}$$

Since g_i equals -1 or 1 for the patients in respectively the low risk or high risk group, the hazard of each risk group is given by:

$$h_{\text{High}}(t) = h_0(t) \exp(\beta) \tag{9.2}$$

$$h_{\text{Low}}(t) = h_0(t) \exp(-\beta) \tag{9.3}$$

The hazard ratio HR is the ratio between the hazard of the two risk groups.

$$HR = \frac{h_{\text{High}}(t)}{h_{\text{Low}}(t)} \tag{9.4}$$

$$= \frac{h_0(t)\exp(\beta)}{h_0(t)\exp(-\beta)} = \exp(\beta)^2 \tag{9.5}$$

With the proportional hazard assumptions [30], the baseline hazard $h_0(t)$ vanishes in the estimation of the hazard ratio. The hazard ratio is computed as the solution of fitting a Cox model with only one covariate g_i and one parameter β. This parameter β maximizes the partial likelihood of the Cox model:

$$L(\beta) = \prod_{i|\delta_i=1} \frac{\exp(\beta g_i)}{\sum_{k\in R(t_i)} \exp(\beta g_k)} \tag{9.6}$$

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta g_i)}{\sum_{k\in R(t_i)} \exp(\beta g_k)} \right]^{\delta_i} \tag{9.7}$$

where $\delta_i = 1$ if patient i experienced the event ($\delta_i = 0$, otherwise) and $R(t_i)$ denotes the set of patients still at risk at time t_i.

It is worth noting that this likelihood is only exact when there are no ties in the event times. This simplification, commonly used in survival softwares, is a good approximationn when there are not too many tied events, and is known as the Breslow approximation [21].

This likelihood (9.7) is known as the partial likelihood (partial in constrast to a full likelihood which takes into account the baseline hazard function not considered here). The maximization of this partial likelihood gives an estimate $\hat{\beta}$ of the Cox model parameter β. Cox shows that maximum partial likelihood estimates are consistent and asymptotically normal [31]. With an increasing number of patients, the estimates converge and are asymptotically unbiased. In many realistic situations, these estimators are also asymptotically fully efficient [40]. This means that with a sufficient number of patients, the variances of the estimates will not be much larger than if we were using the full likelihood. The estimate $\hat{\beta}$ has most of the good properties of a maximum likelihood estimator. Standard MLE inference methods can thus be used on the maximum partial likelihood estimators of the Cox model [31].

In particular, the variance of $\hat{\beta}$ is the inverse of the Fisher information [32] and can be estimated through the second derivative of the log-likelihood with respect to β estimated in $\hat{\beta}$:

$$var(\hat{\beta}) \approx -\left(\frac{d^2 logL(\beta)}{d\beta^2}\right)^{-1}\Bigg|_{\beta=\hat{\beta}} \tag{9.8}$$

9.2.1 Derivatives of the Cox partial log-likelihood

The likelihood of the Cox model is given by

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta g_i)}{\sum_{k\in R(t_i)} \exp(\beta g_k)}\right]^{\delta_i} \tag{9.9}$$

The log-likelihood, its first and second derivatives are given by

$$l(\beta) = \sum_{i=1}^{n} \delta_i \beta g_i - \sum_{i=1}^{n} \delta_i \log\left(\sum_{k\in R(t_i)} \exp(\beta g_k)\right) \tag{9.10}$$

$$\frac{dl(\beta)}{d\beta} = \sum_{i=1}^{n} \delta_i g_i - \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k\in R(t_i)} g_k \exp(\beta g_k)\right)}{\left(\sum_{k\in R(t_i)} \exp(\beta g_k)\right)} \tag{9.11}$$

$$\frac{d^2 l(\beta)}{d\beta^2} = -\sum_{i=1}^{n} \delta_i \left[\frac{\left(\sum_{k\in R(t_i)} g_k^2 \exp(\beta g_k)\right)\left(\sum_{k\in R(t_i)} \exp(\beta g_k)\right) - \left(\sum_{k\in R(t_i)} g_k \exp(\beta g_k)\right)^2}{\left(\sum_{k\in R(t_i)} \exp(\beta g_k)\right)^2}\right] \tag{9.12}$$

The first derivative is known as the efficient score of β and is denoted $u(\beta)$ [28]. The negative of the second derivative of the log-likelihood function is called the observed information function and is denoted $i(\beta)$.

9.2.2 P-value of the hazard ratio

To compute the p-value of the hazard ratio, we can compute the **score test statistic**:

$$\frac{\{u(0)\}^2}{i(0)} \tag{9.13}$$

This statistic has an asymptotic chi-squared distribution with 1 d.f. under the null hypothesis that $\beta = 0$. The values $u(0)$ and $i(0)$ can be computed from equations (9.11) and (9.12).

$$u(\beta) = \sum_{i=1}^{n} \delta_i g_i - \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k \exp(\beta g_k)\right)}{\left(\sum_{k \in R(t_i)} \exp(\beta g_k)\right)} \tag{9.14}$$

$$u(0) = \sum_{i=1}^{n} \delta_i g_i - \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k\right)}{\left(\sum_{k \in R(t_i)} 1\right)} \tag{9.15}$$

$$= \sum_{i=1}^{n} \delta_i g_i - \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k\right)}{|R(t_i)|} \tag{9.16}$$

$$i(\beta) = \sum_{i=1}^{n} \delta_i \left[\frac{\begin{array}{r}\left(\sum_{k \in R(t_i)} g_k^2 \exp(\beta g_k)\right)\left(\sum_{k \in R(t_i)} \exp(\beta g_k)\right) \\ - \left(\sum_{k \in R(t_i)} g_k \exp(\beta g_k)\right)^2\end{array}}{\left(\sum_{k \in R(t_i)} \exp(\beta g_k)\right)^2} \right] \tag{9.17}$$

$$i(0) = \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k^2\right)\left(\sum_{k \in R(t_i)} 1\right) - \left(\sum_{k \in R(t_i)} g_k\right)^2}{\left(\sum_{k \in R(t_i)} 1\right)^2} \tag{9.18}$$

$$= \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k^2\right)(|R(t_i)|) - \left(\sum_{k \in R(t_i)} g_k\right)^2}{|R(t_i)|^2} \tag{9.19}$$

The **score test statistic** $\frac{\{u(0)\}^2}{i(0)}$ is the logrank statistic when there is no tied event and when two risk groups are defined with $g_i \in \{0, 1\}$ [28].

9.3 Balanced hazard ratio

The balanced hazard ratio (BHR) computes the hazard ratio between three curves: the survival curves of the high and low risk groups (as for HR) and a third global survival curve over all patients. Each sample is now considered as a member of 2 groups: its actual risk group ($g_i = -1$, for low risk, or $g_i = 1$, for high risk) and the global risk group ($g_i = 0$) for all patients. Such a global risk group represents the hazard (or survival

time) over the whole population of patients and one measures now how much each specific risk group departs from the global curve. The hazard function is now defined over those 3 groups:

$$h_i(t) = h_0(t) \exp(\beta g_i), \text{with } g_i = -1, 0, \text{ or } 1 \tag{9.20}$$

The quantity $h_0(t)$ represents here the hazard of the whole population, $h_0(t)/\exp(\beta)$ the hazard of the low risk group and $h_0(t)\exp(\beta)$ the hazard of the high risk group. The balanced hazard ratio, $BHR = \exp(\beta)$ is simply the multiplicative factor to get the hazard of the high risk from the global hazard or from the low risk to the global one. The hazard ratio between the hazard of the two risk groups (computed with the balanced hazard ratio) is here:

$$\frac{h_{\text{High}}(t)}{h_{\text{Low}}(t)} = \frac{h_0(t)\exp(\beta)}{h_0(t)\exp(-\beta)} = \exp(\beta)^2 = BHR^2 \tag{9.21}$$

However, note that if BHR^2 has a similar interpretation, it is not equal to the HR (as computed in section 9.2). Indeed, the β value is found by fitting a partial likelihood according to the 3 curves (see below).

The estimation $\hat{\beta}$ of the β value from the balanced hazard ratio (see equation (9.20)) is computed through the maximization of a partial likelihood, similarly to the original HR. For the BHR, the partial likelihood is slightly modified to include the global survival without actually duplicating the patients. The partial likelihood for the balanced hazard ratio (with the standard Breslow approximation for ties [28]) is:

$$L_{BHR}(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta g_i)}{\left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1 \right)^2} \right]^{\delta_i} \tag{9.22}$$

We can compute the BHR log-likelihood and its derivatives similarly to the HR log-likelihood.

$$l(\beta) = \sum_{i=1}^{n} \delta_i \beta g_i - 2 \sum_{i=1}^{n} \delta_i \log \left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1 \right) \tag{9.23}$$

$$\frac{dl(\beta)}{d\beta} = \sum_{i=1}^{n} \delta_i g_i - 2 \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k \exp(\beta g_k) \right)}{\left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1 \right)} \tag{9.24}$$

$$\frac{d^2 l(\beta)}{d\beta^2} = -2 \sum_{i=1}^{n} \delta_i \left[\frac{\left(\sum_{k \in R(t_i)} g_k^2 \exp(\beta g_k) \right) \left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1 \right) - \left(\sum_{k \in R(t_i)} g_k \exp(\beta g_k) \right)^2}{\left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1 \right)^2} \right] \tag{9.25}$$

Using the same results as for the hazard ratio, the BHR estimate $\hat{\beta}$ has most of the good properties of a maximum likelihood estimator [31].

9.3.1 P-value of the balanced hazard ratio

Usually duplicating patients in a statistical test will decrease the p-value of this test. In the BHR however, the whole set of patients is not truly duplicated but is rather used as a form of normalization. To observe the effect of this normalization, we compute the **score test statistic** which is used to compute the BHR p-value:

$$\frac{\{u(0)\}^2}{i(0)} \tag{9.26}$$

This statistic has an asymptotic chi-squared distribution with 1 d.f. under the null hypothesis that $\beta = 0$. The values $u(0)$ and $i(0)$ can be computed from equations (9.24) and (9.25).

$$u(\beta) = \sum_{i=1}^{n} \delta_i g_i - 2 \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k \exp(\beta g_k)\right)}{\left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1\right)} \tag{9.27}$$

$$u(0) = \sum_{i=1}^{n} \delta_i g_i - 2 \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k\right)}{\left(\sum_{k \in R(t_i)} 2\right)} \tag{9.28}$$

$$= \sum_{i=1}^{n} \delta_i g_i - 2 \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k\right)}{2|R(t_i)|} \tag{9.29}$$

$$= \sum_{i=1}^{n} \delta_i g_i - \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k\right)}{|R(t_i)|} \tag{9.30}$$

We observe that $u(0)$ computed with the BHR is equivalent to $u(0)$ computed with the hazard ratio (see section 9.2.2).

$$i(\beta) = 2 \sum_{i=1}^{n} \delta_i \left[\frac{\begin{array}{r}\left(\sum_{k \in R(t_i)} g_k^2 \exp(\beta g_k)\right)\left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1\right) \\ - \left(\sum_{k \in R(t_i)} g_k \exp(\beta g_k)\right)^2\end{array}}{\left(\sum_{k \in R(t_i)} \exp(\beta g_k) + 1\right)^2} \right] \tag{9.31}$$

$$i(0) = 2 \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k^2\right)\left(\sum_{k \in R(t_i)} 2\right) - \left(\sum_{k \in R(t_i)} g_k\right)^2}{\left(\sum_{k \in R(t_i)} 2\right)^2} \tag{9.32}$$

$$= 2 \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k^2\right)(2|R(t_i)|) - \left(\sum_{k \in R(t_i)} g_k\right)^2}{4|R(t_i)|^2} \tag{9.33}$$

$$= \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k^2\right)(|R(t_i)|) - \frac{1}{2}\left(\sum_{k \in R(t_i)} g_k\right)^2}{|R(t_i)|^2} \tag{9.34}$$

The value $i(0)$ computed with the balanced hazard ratio is very similar to the one computed with the hazard ratio (see section 9.2.2). The only difference is the $\frac{1}{2}$ factor which is present here. Since the value $\left(\sum_{k \in R(t_i)} g_k\right)^2$ is positive, the fisher (observed) information function $i(0)$ should be greater for the balanced hazard ratio. The BHR **score**

test statistic $\frac{\{u(0)\}^2}{i(0)}$ is thus smaller while having the same asymptotic chi-squared distribution with 1 d.f., under the null hypothesis that $\beta = 0$. The p-value computed with a balanced hazard ratio is thus conservative with respect to a hazard ratio.

As the groups are defined with $g_i \in \{-1, 1\}$, we can further simplify equation (9.34).

$$i(\beta) = \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} g_k^2\right) (|R(t_i)|) - \frac{1}{2}\left(\sum_{k \in R(t_i)} g_k\right)^2}{|R(t_i)|^2} \tag{9.35}$$

$$= \sum_{i=1}^{n} \delta_i \frac{\left(\sum_{k \in R(t_i)} 1\right) (|R(t_i)|) - \frac{1}{2}\left(\sum_{k \in R(t_i)} g_k\right)^2}{|R(t_i)|^2} \tag{9.36}$$

$$= \sum_{i=1}^{n} \delta_i \frac{|R(t_i)|^2 - \frac{1}{2}\left(\sum_{k \in R(t_i)} g_k\right)^2}{|R(t_i)|^2} \tag{9.37}$$

We can observed that this additional factor ($\frac{1}{2}$) will have no impact when the $\sum_{k \in R(t_i)} g_k$ are close to zero. Those values are close to zero when the class are well balanced. This is consistent with the empirical observation that HR and BHR tend to the same values in this special case.

9.4 Conclusion

In chapter 8, we presented the balanced hazard ratio for risk groups evaluation from survival data. This chapter shows that in the two risk groups setting the BHR p-value is conservative with respect to the HR p-value. Similarly as what was observed with the BHR and HR metrics, their p-values are closer when the risk groups are well balanced. Our futur works include extensions of these results to a number of risk groups larger than 2.

Part III

Hypoxia Signatures for Cancer Prognosis

Chapter 10

Hypoxia Signatures for Cancer Prognosis

In this chapter, we investigated the use of hypoxia-related gene signatures as prognostic biomarkers of cancer progression. In particular, we developed two signatures from the transcriptomic analysis of tumor cell lines exposed to cycling and continuous hypoxia, the CycHyp and ContHyp signatures, respectively. Clinical data sets were then used to assess their prognostic performance.

Unlike the rest of this machine learning thesis, this chapter is more focused on biomedical findings than on methodological issues. This work resulted from a collaboration between the teams of Olivier Feron and Pierre Dupont, my two thesis supervisors.

This work was however the trigger for most methodological contributions developed afterwards, such as the Coxlogit model and Balanced hazard ratio. For the thesis, we actually re-analysed the results presented in our paper [15] in the light of the balanced hazard ratio (chapter 8). Results previously reported in terms of hazard ratio, C-index and BCR remained fully concordant with those obtained using the balanced hazard ratio.

- Boidot, R., Branders, S., Helleputte, T., Rubio, L. I., Dupont, P., and Feron, O. (2014). A generic cycling hypoxia-derived prognostic gene signature: application to breast cancer profiling. *Oncotarget*, **5**(16)

- Feron, O., Boidot, R., Branders, S., Dupont, P., and Helleputte, T. (2015). Signature of cycling hypoxia and use thereof for the prognosis of cancer. WO Patent App. PCT/EP2014/066,643

10.1 Motivations

Hypoxia is nowadays described as a hallmark of tumors [115, 10]. Tumor angiogenesis and glycolytic metabolism are two extensively studied responses of cancer cells to a deficit in oxygen [115]. The building of new blood vessels to bring O_2 and the respiration-independent metabolism to survive under low O_2 are actually complementary responses of tumors to hypoxia [115, 10]. These somehow opposite modes of adaptation account for local and temporal heterogeneities in tumor O_2 distribution. The terms "intermittent hypoxia" or "cycling hypoxia" were settled to describe this phenomenon of fluctuating hypoxia in tumors [22, 39]. As a corollary, the extent of cycling hypoxia reflects tumor plasticity and thus measures the capacity of tumor cells to survive and proliferate in a hostile environment [22].

Although the existence of cycles of hypoxia and/or ischemia was demonstrated in mouse, canine and human tumors [38, 146], technologies aiming to routinely measure tumor O_2 fluctuations in the clinics are not (yet) available despite important progresses in the in vivo imaging of hypoxia [6, 5, 91, 26, 80]. In the absence of readily accessible monitoring strategies, the analysis of the transcriptome associated with this phenomenon could represent a prognostic biomarker of cancer progression. Indeed, although mutations and defects in tumor suppressor genes directly influence the whole genetic profile of a given tumor cell clone, cycling hypoxia could be envisioned as a supra-oncogenic phenomenon influencing gene expression [22]. In other words, independently of the genetic background of tumor cells, cycling hypoxia has the potential to lead to common alterations in the expression of some transcripts, and thus to a possible clinically exploitable signature.

Clinical data sets derived from breast cancer patients could be used to evaluate the performance of such cycling hypoxia-related gene signature. The clinical and genetic heterogeneities of this disease and the very large panel of data sets available represent indeed good opportunities to evaluate new prognostic gene expression signatures [105]. Whole genome analysis already provided several molecular classifications for breast cancer beyond standard clinicopathologic variables [105, 100, 141, 98, 123, 143, 125, 122, 85, 119]. The latter include tumor size, presence of lymph node metastasis and histological grades [102] but also encompass three predictive markers of response, namely expression of oestrogen (ER), progesterone (PR) and HER2 receptors [105]. Treatment guidelines are nowadays still largely based on algorithms integrating these informations such as the Notthingham Prognostic Index [102, 54] or Adjuvant! Online [103]. Accordingly, for early-stage breast cancer, adjuvant

chemotherapy is recommended for most patients with ER-negative or HER2-positive tumors [100, 42, 41, 70]. The challenge actually resides in selecting patients with ER-positive HER2-negative disease who could benefit from chemotherapy.

The aim of this study was:

- To extract a gene signature of cycling hypoxia, the CycHyp signature.
- To better understand the differences between cycling and continuous hypoxia.
- To build a prognostic model with CycHyp in order to yield better breast cancer prognostication, in particular for ER-positive HER2-negative patients.
- To confirm the link between hypoxia and cancer prognosis.

10.2 Materials and methods

The general protocol of this study, summarized in figure 10.1, can be decomposed in three main parts.

First, twenty cell lines derived from various human tumors were exposed to three controlled hypoxia conditions: normoxia, continuous hypoxia and cycling hypoxia. Section 10.2.1 presents the different cell lines used and how the three conditions where reproduced.

The transcriptome associated with these conditions was then analysed to identify two hypoxia-related gene signatures (section 10.2.2). The CycHyp signature (resp. ContHyp) is formed with the most differentiated probesets (genes) between normoxia and cycling (resp. continuous) hypoxia.

The sections 10.2.4 and 10.2.5 present respectively the data sets and the construction of the models used to validated the signatures on public breast cancer data sets.

10.2.1 Cell lines data

Twenty cell lines derived from various human tumors and characterized by a large variety of distinct genetic anomalies (see table A.1 in appendix) were submitted to cycling hypoxia (CycHyp), i.e. 24 cycles of 30 min incubation under normoxia and 30 min incubation under hypoxic (1% O2) conditions to reproduce tumor hypoxic fluctuations, as previously reported [38, 37]. At the end of some hypoxic periods (after

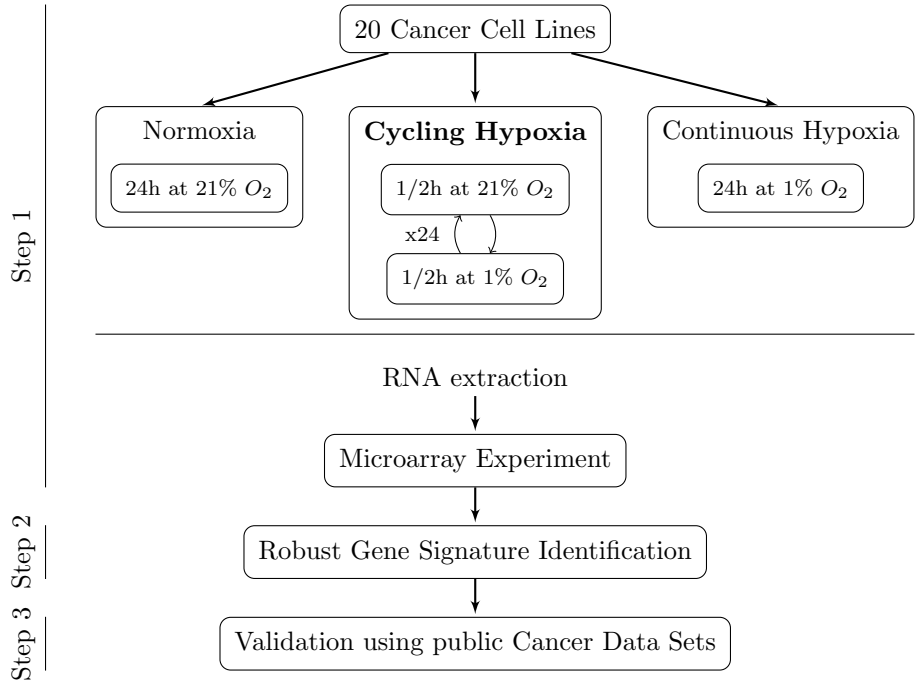


Figure 10.1: General scheme of the study: from cell lines to signature validation.

1, 2, 3 and 24 cycles), the induction of the transcription factor HIF-1α was confirmed by immunoblotting. Also, we verified that the HIF-1α immunoblot signal was absent after 30 min incubation under normoxia (post-hypoxia). These results validated that the imposed changes in the atmospheric pO2 were sensed by the cells bathing in their culture medium. We also considered control conditions of 24 h continuous exposure of tumor cells to either 21% O_2 (Normoxia) or 1% O_2 (ContHyp). For each culture condition, cells were immediately snap-frozen at the end of the last incubation period. mRNA extracts from each tumor cell cultured under the three above conditions (normoxia, cycling hypoxia and continuous hypoxia) were analysed by hybridization on Human Gene 1.0 ST Affymetrix microarrays (GEO access number: GSE42416).

10.2.2 Identification of the CycHyp signature

Gene expression profiles of each cell type under normoxia, continuous hypoxia and cycling hypoxia were produced to identify two gene signatures: CycHyp and ConHyp. The CycHyp signature is made of differentially expressed genes between cycling hypoxia and normoxia. The ContHyp

signature is made of differentially expressed genes between continuous hypoxia and normoxia.

A prefiltering is performed first to remove non-annotated and control probesets. This preprocessing step simplifies the signature interpretation and the signature transfer to other affymetrix microarray platforms.

The methodology used to identify the signatures is an ensemble feature selection similar to [1]. It relies on a resampling mechanism to repetitively look at the data from complementary viewpoints and to build a signature robust to small data perturbations. For every resampling experiment, a subset of 90% of the samples is chosen uniformly at random without replacement and forms a training set. Differentially expressed probesets are assessed on each subset according to a paired t-test to discriminate between both conditions. FDR corrected p-values are reported. The top 100 probesets with the smallest corrected p-values averaged over 200 resamplings form the final signature.

The t-test was chosen as it is both simple and computationally efficient. More complex alternatives exist such as SAM [135] and limma [118]. They are both moderated t-statistics similar to the t-test except that the variance is moderated across genes. However, the three methods have very similar results in terms of gene lists produced [74, 73]. Moreover, we use the methodology proposed by Boulesteix *et al.* [16] to confirm that the t-test rankings are not affected by outliers or extreme values (see appendix A.2).

For each resampling experiment, the 10 % remaining data form a small independent test set used to estimate the discrimination capability between both conditions. More specifically, a linear SVM (presented in section 3.2.2) is estimated using the 100 probesets with the lowest p-values on the 90 % training and its classification accuracy between normoxia and cycling (resp. continuous) hypoxia is evaluated on the validation set. The average accuracy is reported over 200 resamplings. This evaluation protocol is summarized hereafter.

1. Remove non-annotated and control probesets
2. Repeat 200 times :
   (a) Random sampling of the data (20 samples in each condition): 90% train, 10% test
   (b) For each probeset, compute an FDR corrected p-value from a paired t-test (on train)
   (c) Estimate on train a linear SVM using the 100 probesets with lowest p-values

(d) Compute the SVM classification accuracy on the test

3. Average the FDR corrected p-values and accuracy over the 200 resamplings

The 100 probesets with the lowest p-values, averaged over 200 resamplings, formed the final CycHyp (resp. ContHyp) signature discriminating between normoxia and cycling (resp. continuous) hypoxia. These two signatures are reported in table A.2 and table A.4.

We use this resampling mechanism to produced robust gene signatures and to have estimates of their predictive power. Doing so, we avoid an optimistic selection bias that would result from considering the same data to identify signatures and to estimate their predictive power. Yet, those results do not represent the predictive power of the CycHyp and ContHyp signatures themselves. Indeed, those signatures are eventually estimated by averaging results over all resamplings, leaving no independent *cell lines* data to assess their predictive performances. We do not consider this to be problematic since the predictive power of those signatures is evaluated anyway on a much harder task and independent data: a risk group prognosis from human breast cancer samples.

10.2.3 Transfer of the signatures across microarray technologies

The CycHyp and ContHyp signatures were identified on cell lines using the Affymetrix HGU1.0ST microarray technology (see sections 10.2.1 and 10.2.2). Assessing the prognostic value of those signatures on publicly available breast cancer data requires to transfer those signatures to a former generation of Affymetrix chips.

The CycHyp and ContHyp signatures were transferred from the Affymetrix HGU1.0ST to HGU133a using the ENTREZ gene id. The CycHyp (and ContHyp) HGU133a signature is defined as all HGU133a probesets with an ENTREZ gene id present in the original HGU1.0ST signature of 100 probesets. The HGU133a CyHyp signature contains 87 probesets (123 for ContHyp). These two signatures are presented in tables A.3 and A.5.

10.2.4 Breast data sets

Table 10.1 presents the 9 data sets of primary breast cancer used to estimate the prognostic potential of the CycHyp and the ContHyp signatures.

	Data set	Technology	Nb patients	source GEO
1	VDX [143, 95]	HGU133a	344	GSE2034 GSE5327
2	TBG [35]	HGU133a	198	GSE7390
3	UNT [124]	HGU133a	189	GSE2990
4	MAINZ [113]	HGU133a	200	GSE11121
5	UPP [94]	HGU133a	251	GSE3494
6	LUM [87]	HGU133plus2	414	GSE6532
7	BRC [76]	HGU133plus2	327	GSE20685
8	IPC [111]	HGU133plus2	266	GSE21653
9	MDA5 [130]	HGU133plus2	298	GSE17705

Table 10.1: Primary breast cancer data sets used here.

All breast cancer expression data were summarized with MAS5 and represented in log2 scale (except for GSE6532 already summarized with RMA). The HGU133plus2 data sets were reduced to their HGU133a probesets. Three breast cancer subtypes (ER+/HER2-, ER-/HER2- and HER2+) were identified with the genefu R package [61]. Disease-free survival was used as the survival endpoint. The data from all patients were censored at 10 years to have comparable follow-up times across clinical studies [60].

10.2.5 Prognostic model construction

The CycHyp model was estimated on the VDX data set (on the ER+/-HER2-, node-negative and untreated population) and assessed on the 8 remaining datasets listed in the table 10.1. We used the VDX dataset as training because of its large number of node negative untreated patients. Each data set is reduced to the population of interest and then normalized (scaled and centered) separately (z score). A risk score for each patient was computed from a penalized Cox proportional hazards model (see section 3.3.1) implemented in the Penalized R package [55]. The parameters of the elastic net penalty were learned on the training set by cross-validation. The final L_1 and L_2 penalties are fixed to 10 and 16, respectively.

To have comparable risk scores accross the data sets, the CycHyp score was scaled per data set such that the 2.5% and 97.5% percentiles are set to -1 and 1 respectively. This scaling, used in [61], is robust to outliers.

A threshold on the risk scores is used to classify patients between high and low risk groups. This threshold is computed from the ROC curves (time dependent ROC curves at 5 years) to maximize the average between sensitivity and specificity. This average is also named balanced

classification rate (BCR) (see section 4.2.5).

The very same CycHyp model was used to predict the clinical outcome on the validation sets (8 datasets) while considering various subpopulations of interest (all patients or the ER+/HER-, node- and ER+/-HER- population). This procedure has the benefit to compare the prediction capabilities of always the same model for various populations of patients. Additional experiments (not detailed) showed that results are essentially unchanged if one estimates on VDX a different model for each population of interest.

10.3 Results

10.3.1 The CycHyp signature

The CycHyp and ContHyp signatures were identified using a resampling protocol described in section 10.2.2. The 100 probesets with the lowest p-values, averaged over 200 resamplings, formed the final CycHyp (resp. ContHyp) signature discriminating between normoxia and cycling (resp. continuous) hypoxia. These two signatures are reported in table A.2 and table A.4.

The resampling protocol used in the feature selection allows us to compute an estimate of the predictive performance of the CycHyp signature. The accuracy to classify independent samples is on average 97.46% for cycling hypoxia versus Normoxia and 94.33% for continuous hypoxia versus normoxia. The heatmaps made with the 100 probe sets of the CycHyp signature confirmed its excellent potential of discrimination between cycling hypoxia and either normoxia (figure 10.2) or continuous hypoxia (figure 10.3). Each row in the heatmaps corresponds to a particular cell line in either normoxia, cycling hypoxia or continuous hypoxia. These cell lines are identified by a number from 1 to 20. The numbers can be used to access more information about each cell lines in table A.1 in appendix. Similar results are obtained with the ContHyp signature (heatmaps A.4 and A.5 in appendix).

Moreover, Gene Set Enrichment Analysis (GSEA) [129] indicated that when considering differentially expressed probesets (after FDR correction), only 2 gene sets were significantly enriched in the CycHyp signature (table A.6) whereas we identified 52 gene sets enriched in the ContHyp signature, including 17 directly related to hypoxia (table A.7).

Also, when using the MSigDB molecular signature database referring to hypoxia or HIF (www.broadinstitute.org), we found 13 hypoxia gene sets sharing, on average, only 1.4 gene with CycHyp (table A.8 in appendix) whereas 44 hypoxia gene sets showed overlap with ContHyp

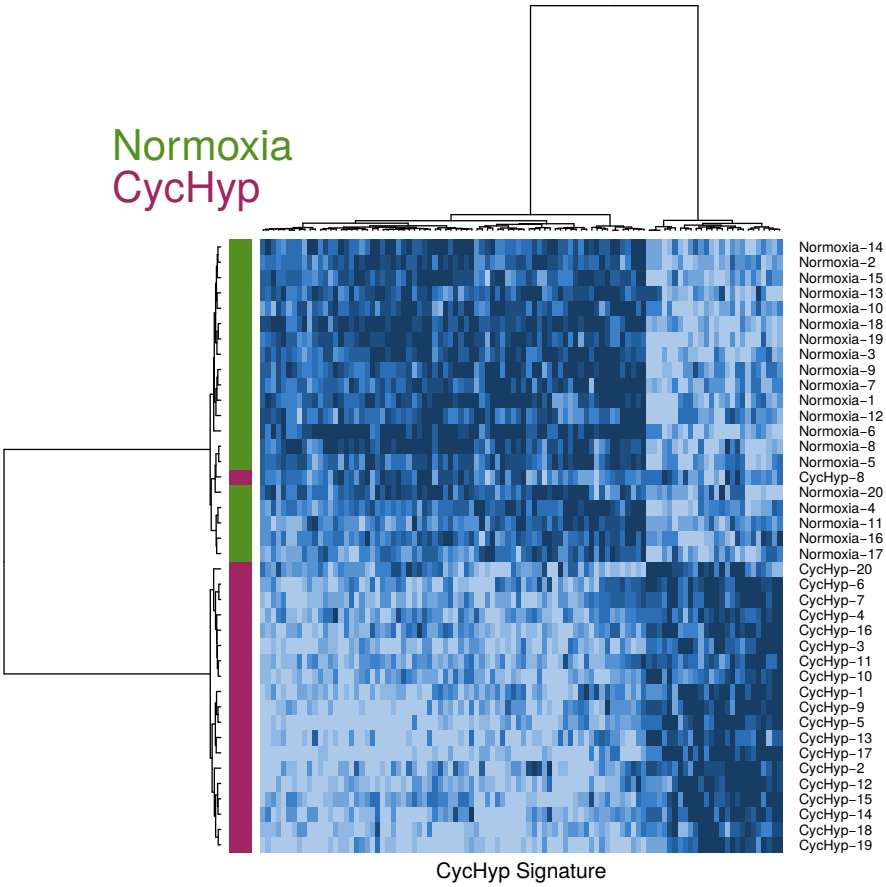


Figure 10.2: Heatmap depicting the transcripts from the CycHyp signature either underexpressed (light blue) or overexpressed (dark blue) (centered to median values). Each column corresponds to a specific human Gene 1.0 ST probeset ; each line represents a specific cell line either maintained under normoxia (green) or exposed to cycling hypoxia (purple). The numbers (from 1 to 20) can be used to access information about specific cell lines in table A.1 in appendix.

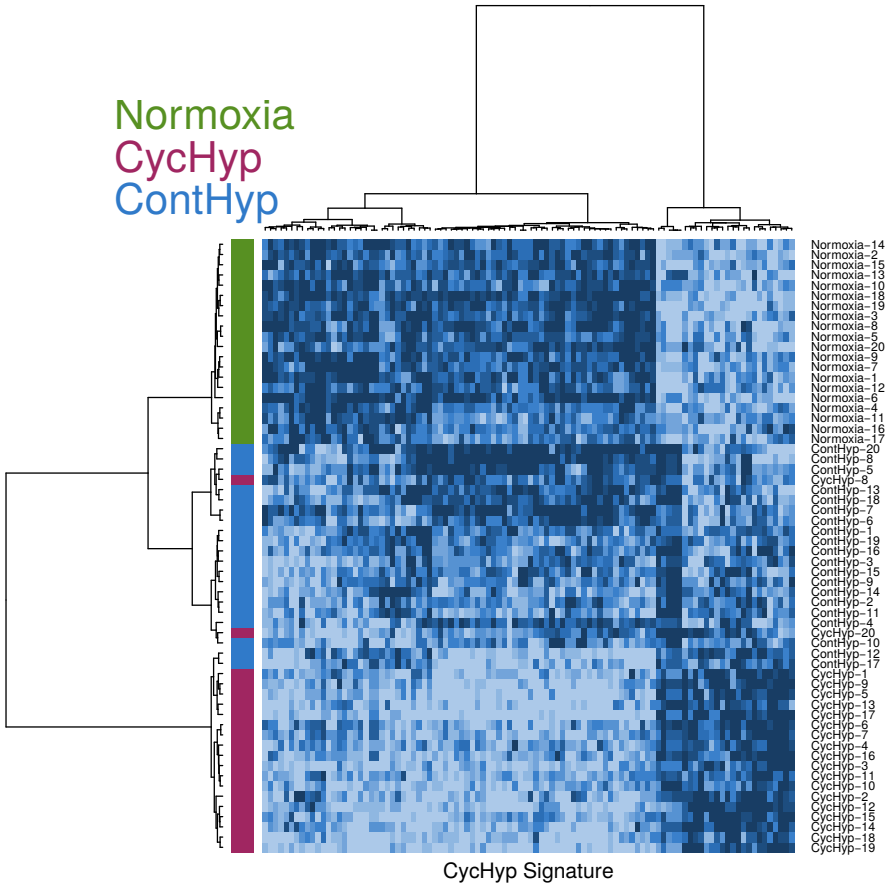


Figure 10.3: Heatmap depicting the transcripts from the CycHyp signature either underexpressed (light blue) or overexpressed (dark blue) (centered to median values). Each column corresponds to a specific human Gene 1.0 ST probeset ; each line represents a specific cell line either maintained under normoxia (green), exposed to continuous hypoxia (blue) or exposed to cycling hypoxia (purple). The numbers (from 1 to 20) can be used to access information about specific cell lines in table A.1 in appendix.

with an average of 6.6 (1-27) common genes (table A.9 in appendix).

We also compared the CycHyp signature to 13 other hypoxia-derived signatures described by Seigneuric *et al.* [114] and Starmans *et al.* [126]. The CycHyp signature was again far from those signatures with an average of only 1 gene in common. The overlap was larger between ContHyp and those signatures with an average of 6 genes in common (table A.10 in appendix).

Finally, using TFactS [43] to analyse transcription factors regulating expression of genes associated to either signature, HIF-1α was only found as positively associated with the ContHyp signature.

10.3.2 Validation on breast cancer data

To evaluate the prognostic value of the CycHyp signature, we focused on breast cancer because of the very large amounts of well-annotated clinical data sets available and a clearly identified need to discriminate between patients at low and high risks among subgroups determined on the basis of clinicopathologic criteria [105, 100]. Publicly available GEO data sets allowed us to collect information on the survival of 2,150 patients with primary breast cancer (see section 10.2.4).

In order to exploit these data sets, we first transferred the Gene 1.0ST datasets in the HU133 platform (see section 10.2.3). We then used the VDX dataset (GSE2034 and GSE5327) as a reference because of its large number of node negative untreated patients [143]. This training dataset was used to estimate a prognostic multivariate Cox proportional hazard model built on the CycHyp signature (see section 10.2.5 for details). The other eight datasets (table 10.1) were used according to the methodology described by Haibe-Kains and colleagues [60], to assess the prognostic performance of the CycHyp signature on independent samples.

The prognostic performance are reported here in balanced hazard ratio (see chapter 8). The balanced hazard ratio (BHR) is very similar to the hazard ratio but with better properties. With two risk groups, the BHR compare the hazard of the global population of patients with the hazard of each risk groups. It computes the multiplicative increase (resp. decrease) of hazard for the patients at high risk (resp. low risk) with respect to the average hazard of the patients. For the readers unfamiliar with this metric, the squared balanced hazard ratio (BHR^2) has the same scale and can be compared to a standard hazard ratio.

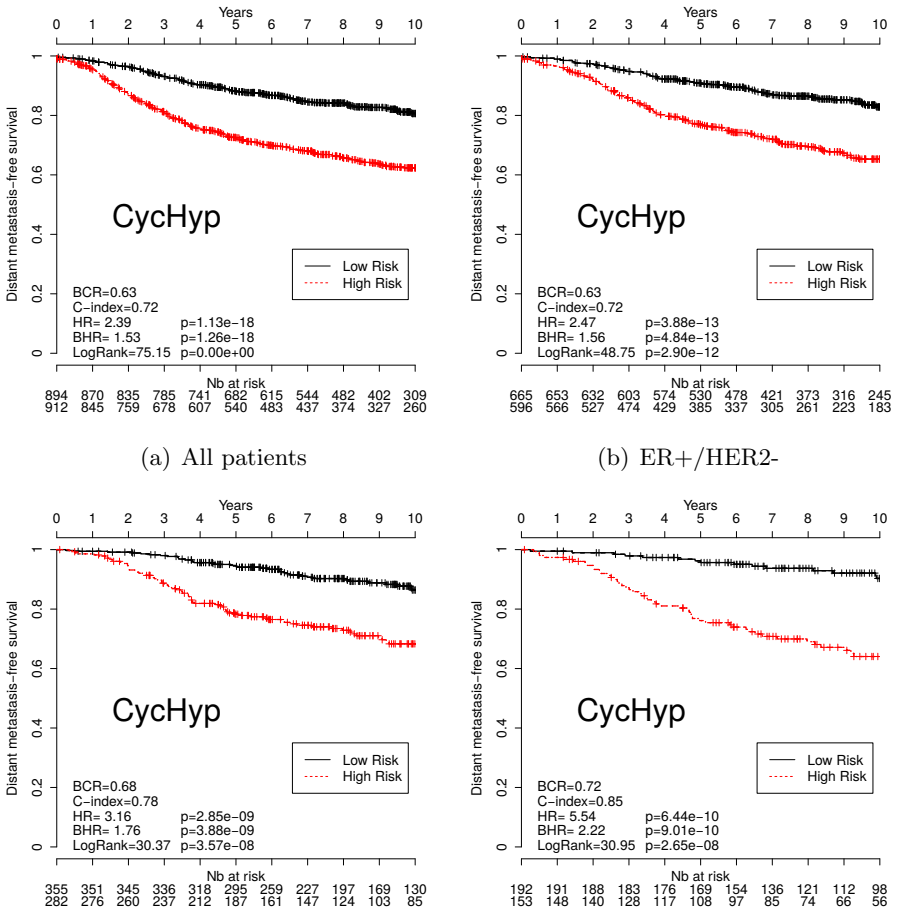
The CycHyp signature predicts clinical outcome in breast cancer patients

The prognostic potential of the CycHyp signature to discriminate between patients at low or high risk was confirmed with a balanced hazard ratio of 1.53 and a p-value=1.26e-18 whathever the treatment and the tumor histology (figure 10.4(a)). We then focused on the ER+/HER2- population which is known to be heterogeneous and thus difficult to treat [105, 100]. The discriminating capacity of the CycHyp signature remained strikingly high in the ER+/HER2- patient populations (BHR=1.56, p-value=4.84e-13, figure 10.4(b)). Finally, among this subpopulation of patients, we considered those with a node negative status (figure 10.4(c)) and among the latter, those who did not receive any treatment (figure 10.4(d)). Balanced hazard ratios rose to 1.76 and 2.22 in these conditions (p-values=3.88e-9 and 9.01e-10, respectively), further supporting the discriminating potential of the CycHyp signature. In particular, the data presented in figure 10.4(d) allowed to exclude any confounding influence of the potential benefit arising from the treatment administered to these patients and thus clearly identified a population of patients who remained inadequately untreated. Other subpopulations of patients were considered in appendix A.8.

Using the same methodology, we examined the prognostic capacity of the ContHyp signature (discriminating between normoxia and continuous hypoxia). The performance of the ContHyp signature was satisfactory on the ER+/HER2- untreated population (BHR=1.61, p-value=1.93e-4, see figure 10.5(a)) but was significantly lower (p-value= 0.012, see figure 10.6) than the CycHyp signature.

The CycHyp signature provides significant additional prognostic information to available multigene assays

To evaluate the performance of the CycHyp signature, we compared it with other well-established prognostic multigene assays for breast cancer, namely Gene70 or Mammaprint [141], Gene76 [143] and Oncotype Dx [98]. Using the same set of ER+ HER2- node-negative untreated patients as used in figure 10.4(d), we could determine the low *vs.* high risk patient stratification according to these signatures (prognostic models). The superior prognostic potential of the CycHyp signature could be captured from the Kaplan Meier curves obtained with the Gene 70, Gene76 and Oncotype DX signatures (compare figure 10.5 with figure 10.4(d)). Balanced hazard ratios confirmed the net advantage of the CycHyp signature with a significantly higher value than the three other metagenes (figure 10.6). Similar results are reported in terms of



(a) All patients

(b) ER+/HER2-

(c) ER+/HER2-, n-

(d) ER+/HER2-, n-, untreated

Figure 10.4: Kaplan-Meier survival curves of patients with primary breast cancer, as determined by using the CycHyp signature.

hazard ratio, concordance index and BCR in our paper [15]. The sensitivity and specificity to discriminate between patients with progressing disease *vs.* disease-free at 5 years are reported in figure 10.6. The sensitivity of the CycHyp signature was above 80% and the specificity was well above the level of the three multigene assays.

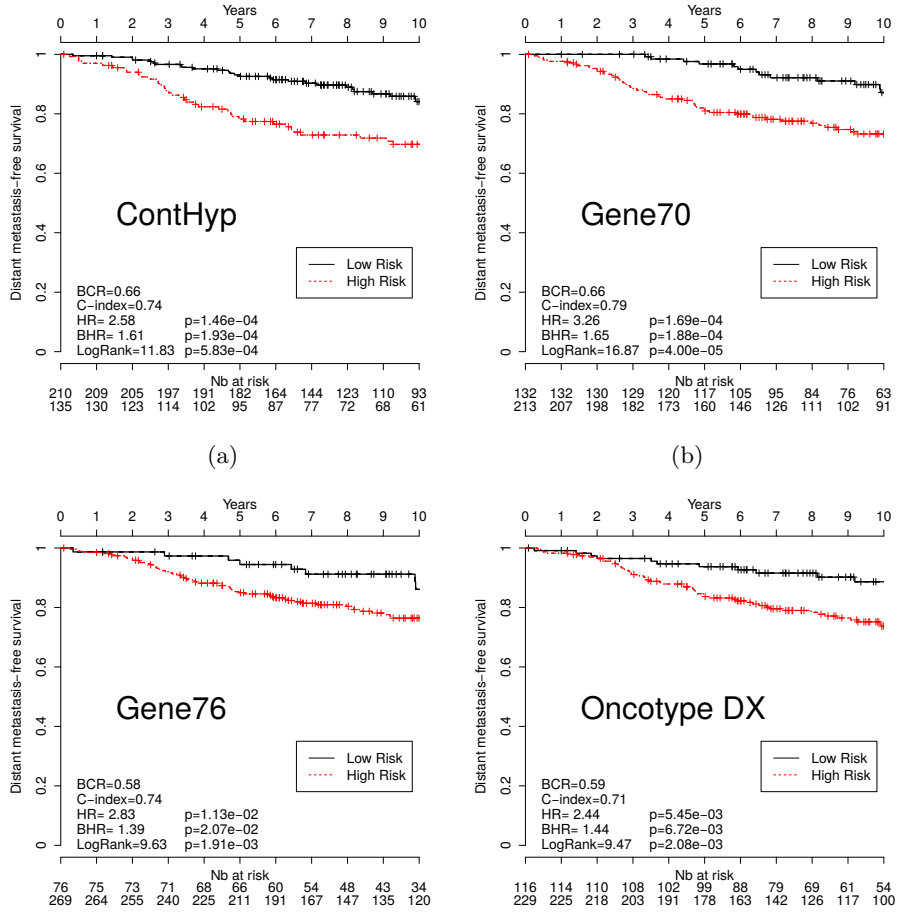


Figure 10.5: Kaplan-Meier survival curves of node-negative, untreated ER+/HER2- patients, as determined by using the ContHyp signature, Gene 70 (Mammaprint), Gene 76 and Oncotype Dx models.

Importantly, to further validate the prognostic significance of the CycHyp signature, a comparison with random gene signatures was performed according to the methodology described by Venet *et al.* [142] and Beck *et al.* [7]. For a fare comparison, the random signature had

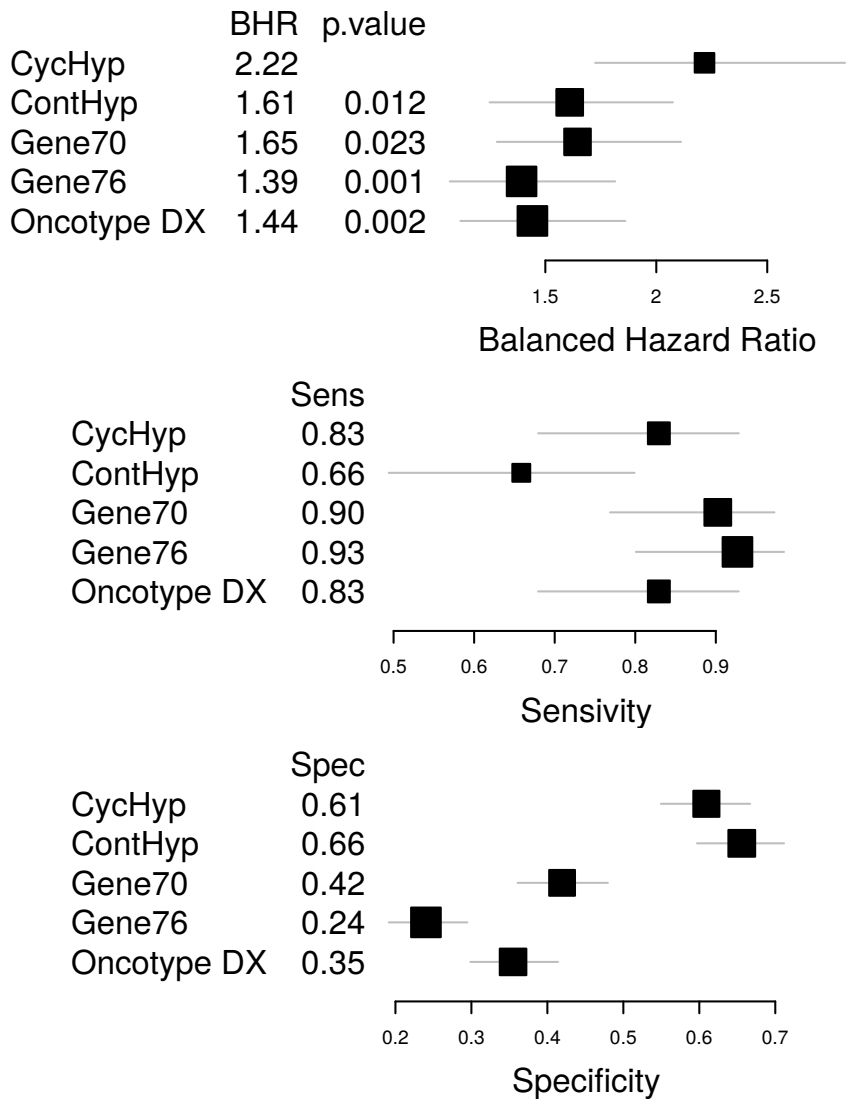


Figure 10.6: Comparison of the prognostic potential of the CycHyp signature *vs.* ContHyp signature, Gene 70 (Mammaprint), Gene 76 and Oncotype Dx models.

the same size as the CycHyp and ContHyp signatures (resp. 87 and 123 probe sets). The data sets and the methodology were also the same as presented in section 10.2.5. Figure 10.7 shows the distribution of the p-values (logrank test in log 10) for 1000 randomly generated signatures together with the p-values of the CycHyp and ContHyp signatures represented with the two red dots. The discrimination between risk groups was significantly higher (p-value < 0.001) with the CycHyp signature as compared to each of the random signatures whereas the ContHyp signature was not significantly better (*vs.* random ones; P=0.141).

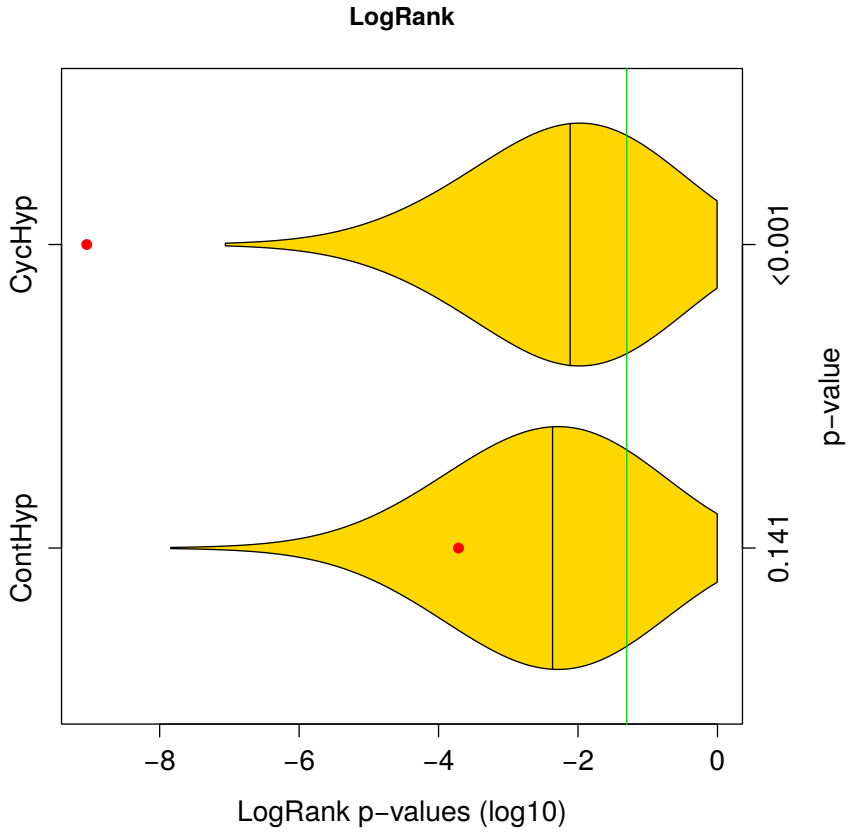


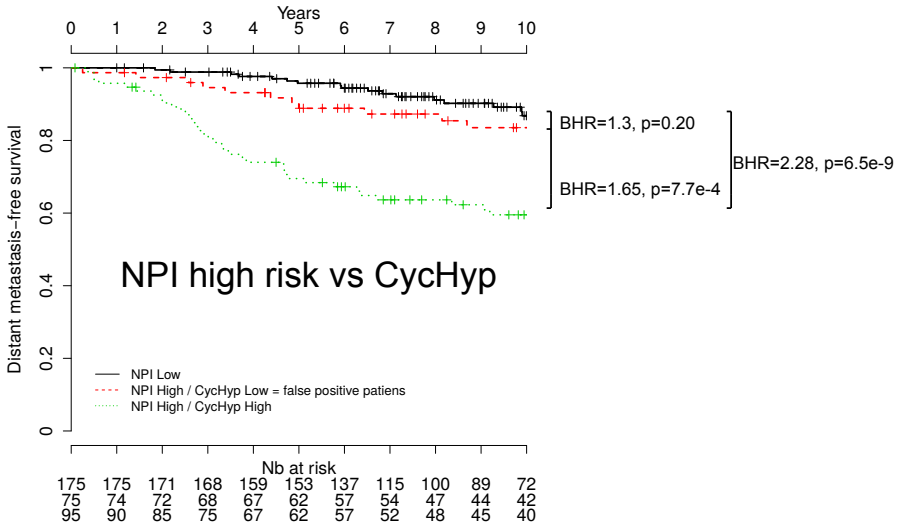
Figure 10.7: Graph represents the power of discrimination in high *vs.* low risk groups (expressed as the logarithm of the p-values of the logrank) of the ContHyp and CycHyp signatures (see red dots) versus 1,000 randomly generated signatures (yellow shapes depicting their distribution).

The CycHyp signature in association with NPI offers a powerful prognostic tool

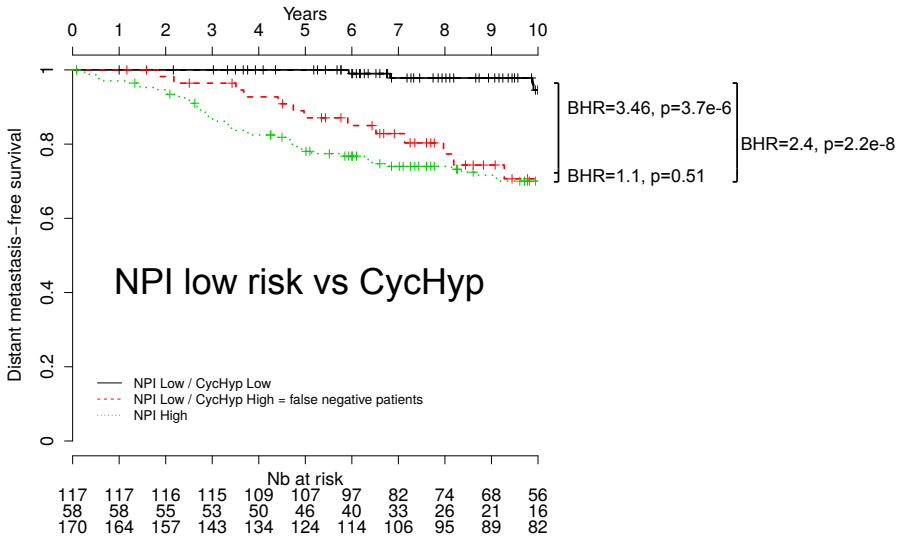
We then aimed to determine whether the CycHyp signature could improve the Nottingham Prognostic Index (NPI) for better predicting the survival of operable breast cancers. The NPI algorithm combines nodal status, tumour size and histological grade and allows to model a continuum of clinical aggressiveness with 3 subsets of patients divided into good, moderate, and poor prognostic groups with 15-year survival [102, 54, 41]. Since very few patients were assigned a poor index, we merged here the moderate and poor indices into a high risk group to facilitate the comparison with the CycHyp signature. We found that by integrating the CycHyp signature, an important proportion of patients could be reclassified to another risk group (figure 10.8). 44.1% of patients classified at high risk using the NPI algorithm were identified at low risk when using the CycHyp signature and were confirmed to be "false positive" since they actually exhibited a profile of survival closer to the low risk NPI patient (figure 10.8(a), BHR=1.3, p-value=0.20). Inversely, using the CycHyp signature, we also identified in the patients at low risk based on the NPI criteria, 33.1% of patients with a risk profile closer to the patients with a negative outcome (figure 10.8(b), BHR=1.1, p-value=0.51). This increased discriminating potential remained highly relevant when considering all patients or patients with a ER+ HER2- status (and among the latter, those with a node negative status or the untreated ones) (see section A.9 in appendix). In appendix A.10, we also assess the performances of the CycHyp signature in each of the six NPI subgroups defined by Blamey *et al.* [13].

10.4 Conclusion and perspectives

This study demonstrates that a gene signature derived from the transcriptomic adaptation of tumor cells to cycling hypoxia is prognostic of breast cancer. The CycHyp signature that we have identified and validated in this study has not only prognostic value independently of molecular risk factors but also provides significant additional prognostic information to clinicopathologic criteria. Clinical outcome of breast cancer patients is nowadays largely based on histological grade and the status of ER, PR, and HER2 receptors [105, 100, 102]. In early breast cancer, a lack of expression of ER (and PR) will almost systematically lead to the administration of adjuvant chemotherapy in addition to locoregional treatment [105, 42, 41]. Also, for patients with a tumor expressing HER2, chemotherapy and/or trastuzumab represents the option



(a)



(b)

Figure 10.8: Kaplan-Meier survival curves of node-negative, untreated ER+/HER2- patients stratified by using the CycHyp signature to detect: (a.) false positive patients among those identified at high risk based on the NPI nomenclature and (b.) false negative patients among those identified at low risk based on the NPI nomenclature.

the most likely to be beneficial based on current clinical knowledge [105]. The impact of chemotherapy is actually more difficult to anticipate for the rest of early-stage breast cancer patients, i.e. those diagnosed with a ER-positive and HER2-negative disease. These patients represent indeed a wide spectrum of different risk profiles: for women with high-risk disease, if chemotherapy is appropriate, others will derive little benefit from it. Our study therefore represents a significant advance for this population of patients, which consists of two third of all breast cancers. We have indeed demonstrated that the CycHyp signature outperforms the existing major prognostic gene expression signatures and offers a unique decision making tool to complement the discrimination of breast cancer patients based on anatomopathologic evaluation.

More generally, the excellent prognostic value of CycHyp confirms the link between cycling hypoxia and cancer aggressiveness [39, 38]. This gives credentials to the phenotypic adaptation of tumors resulting from heterogeneities in blood flow distribution as a trigger of cancer progression [22, 39]. Also, with the recent impetus in the understanding of tumor metabolism [79, 47], it has become obvious that the capacity of a given tumor cell to survive in both aerobic and anaerobic environments represents a critical advantage [145, 121, 14]. Interestingly, our study also documents the higher prognostic value of a transcriptomic signature derived from cycling hypoxia vs. continuous hypoxia. This confirms that although hypoxia is a frequent feature of poor-prognosis tumors and was reported to drive gene signature associated with negative outcome [25, 144, 23, 45], prognostic markers integrating fluctuations in the hypoxic status of tumors (this study) introduce an additional layer of complexity that better fits the in vivo situation.

Whether the CycHyp signature encompasses genes that actively drive cancer progression or reflects a context of metabolic and hypoxic stress favorable to increased mutagenesis and genetic instability [22], warrants further studies. A few hints can however be gleaned from the comparison of the different signatures.

First, the comparison of the CycHyp and ContHyp signatures indicates that the cycling nature of hypoxia leads to specific alterations in mRNA expression since only 11 common transcripts were found in the two gene lists (see symbols ♯ in table A.2). Furthermore, among these 11 genes, most encode for proteins involved in housekeeping functions such as chromatin packaging (HIST1H 1C, 2AC, 4A and 4C) and RNA processing (RPS13 and 28). The only gene common to the two signatures with a known function related to hypoxia is RBX1 or E3 ubiquitin ligase which mediates the ubiquitination and subsequent proteasomal degradation of target proteins [92], including the misfolded proteins known to

accumulate under low pO2. Besides the RBX1 gene, the CycHyp signature does not actually contain genes known to be consistently regulated in response to chronic hypoxia. By contrast, the ContHyp signature contains 14 genes already reported to be overexpressed under low pO2 and even directly under the control of the transcription factor HIF-1α, including those coding for glucose metabolism enzymes (ALDOA, PFKB3, PFKB4, PGK1, PGAM1, GPI) and the angiogenic growth factor VEGFA. This HIF-dependent gene expression program of the ContHyp signature was actually confirmed in the GSEA and MSigBD analyses and was consistent with previously reported hypoxia-driven gene signatures [25, 23, 45]. More generally, these findings position the CycHyp signature far from the conventional hypoxia-derived signatures [114, 126] but instead as a biomarker of a distinct tumor biology process involving adaptation to fluctuations in the tumor microenvironment.

Second, a large amount of transcripts of the CycHyp signature encode for proteins themselves involved in the regulation of transcription. Data mining revealed that more than 18 transcripts of the CycHyp signature are transcription factors/regulators and 13 others are directly involved in RNA processing (see symbols $*$ and § in table A.2 respectively). This represents one third of the genes comprising the CycHyp signature and reflects a major difference with the ContHyp signature. While hypoxia is usually associated with cell cycle arrest and mTOR inhibition, cycling hypoxia may be compatible with a maintained proliferation potential. This is further supported by the suppression of geroconversion (ie, the process leading from proliferative arrest to irreversible senescence) observed in response to hypoxia [81, 82] that offers tumor cells the opportunity to re-enter cell cycle when O_2 is again available. Further studies are needed to compare the evolution of mTOR activity and mTOR-dependent genes (including those encoding for ribosomal proteins) during cycling and continuous hypoxia.

Finally, the in vitro conditions at the origin of the establishment of the CycHyp signature may actually have specific bearing on its robustness and applicability. Indeed, we previously documented that fluctuating oxygen levels could also directly impact endothelial cells within a tumor [90, 33] indicating that non-tumor cells may also contribute to the same transcriptomic adaptation as tumor cells, thereby reinforcing the relevance of the CycHyp signature. Also, although we have used the CycHyp signature as a prognostic biomarker for early-stage breast cancer, this signature was identified by integrating the information arising from tumor cells of various origins and characterized by various oncogenic alterations; the prognostic value of the CycHyp signature in other cancers is currently under investigation in our laboratory.

Altogether, the above findings indicate that the CycHyp signature represents a new generation of prognostic biomarker reflecting a generic environmental condition in tumors that differs from the conventional view of a static, continuous hypoxia occurring in tumors. When applied to breast cancer, the CycHyp signature has a powerful prognostic value independently of molecular risk factors but also offers a unique decision making tool to complement the discrimination of patients based on anatomopathologic evaluation. The CycHyp signature is distinct from conventional hypoxia-related gene signature but also from existing prognostic metagenes, and the rationale behind its discovery supports a potential broad applicability to evaluate cancer patient outcomes.

Part IV

Conclusions and Perspectives

Chapter 11

Conclusion

In this thesis, we focus on prognostic models of cancer progression. We investigate most steps in the creation of new models from the feature selection and the construction of models to the validation of their prognostic potential.

A particular attention was given to the prediction of risk groups through the development of the Coxlogit model and the balanced hazard ratio, respectively presented in chapters 6 and 8. We focus on this aspect of the survival analysis as it is too often overlooked. If the survival analysis is usually viewed as a regression problem (*e.g.* the prediction of a continuous risk score) and can be efficiently tackled as such, the true finality of the survival prediction and prognostic models is to make decisions. For example, clinicians are routinely required to decide whether a specific treatment should be considered for a given patient. Such decision is precisely based on the assignment of this patient to a particular risk group, which is a classification task. Moreover, the challenges of risk group prediction should be tackled as most survival results are reported as such in the medical literature.

11.1 Future works and perspectives

This section presents some perspective and future works that could be done to further improve prognostic models and their validation. Firstly, we list some direct extensions to the models and methods presented in this thesis. We then discuss two important aspects in the estimation of prognostic models: the choice of hyper-parameters and the heterogeneity of cancer. These two points are a personal thought on some methodological problems that still need to be addressed. Why they are

problematic and how addressing them could improve prognostic models in cancer research.

- In this thesis, we proposed the balanced hazard ratio to evaluate survival risk groups. While having a similar interpretation as the original hazard ratio, the BHR penalizes extremely unbalanced risk groups and is much more appropriate to compare predicted risk groups. An online tool and a R package are in development to make the balanced hazard ratio available to any clinician or bio-statistician.

- A Cox model in combination with random non-linear projections can be used to predict the survival of patients from non-linear data. This approach, which is both simple and computationally efficient, could be extended to perform feature selection. Such a non-linear feature selection could be performed while comparing the model parameters and random projections, selecting features with high weight in important projections.

- We have shown the interest of combining two supervisions (risk group labels and survival times) in the Coxlogit model to improve the feature selection and the predictive performances. However, the Coxlogit model is limited to binary risk groups predictions. In many situations, three or more ordered risk groups are available for example the TNM stages that are available in many cancers. To deal with a number of groups larger than 2, we could replace the logistic part of the model by an ordered logistic model.

- If the Coxlogit model could easily be extended to other combinations of generalized linear models, the most promising perspective is its use in transfer learning. One Coxlogit model could indeed be use to link two datasets (or more) without a full or even a partial overlap, *e.g.* one dataset with the survival of some patients and an other with the class labels of other patients. Having a unique model will enforce a consistency between the two tasks in the selection of jointly predictive genes.

- In chapter 8, we discuss the use of the balanced hazard ratio to select a cutoff on risk scores. This cutoff is then used to predict the risk groups. In chapter 6, we used external labels (such as the tumor grade, *etc*) to improve the risk group prediction. Within the Coxlogit model, an implicit cutoff is selected and could be compare to the one selected be the BHR.

11.1.1 Hyper-parameter and cross-validation

Through this thesis, I observed the difficulties of tuning hyper-parameters through internal cross-validations. Examples of hyper-parameters are the L_1 and L_2 regularization parameters (see section 3.1.1) but could also be the size of the gene signature, *etc.* We perform internal cross-validations (cross validations inside the training set) to estimate the effects of each hyper-parameter. This procedure allows us to select a set of hyper-parameters that will be used to train the model on the whole training set. We rely on cross-validation to avoid any optimistic bias on the final performance estimated on the validation set (or on the external cross-validation). Indeed, tuning the parameters to maximize directly the performance on the validation set will overestimate the performances of the model [4].

However, if this cross-validation is needed and important, I observed that it is not a very effective way to tune hyper-parameters. The distribution of the selected parameters is often not far from a uniform random distribution. This variability in the hyper-parameters tend to increase the variance of the final reported performances. It could also lead to the selection of hyper-parameters that are outside the acceptable range, *e.g.* a too strong L_1 regularization which gives a null model.

This observation is particularly true in our domain where the number of samples is very low. This small number of samples is even more reduced while tuning the hyper-parameters in an internal cross-validation. The amount of available samples can be problematic to build a model but also to estimate its performance: How can we reliably estimate a hazard ratio (or a balanced hazard ratio) on 10 samples? We can even show that the balanced classification rate (section 4.1.3) converge to the accuracy (section 4.1.1) when estimated in cross-validation with a small number of samples.

Even if we could have done a proper assessment of the performances of each parameter, we are often at the beginning of the learning curve where a prognostic model changes the most with the number of samples. The optimal set of parameters on 90% of the training data may not be the right one on the full training set. Examples of learning curves are presented in section 6.3.3 on breast cancer data. In these examples, the learning curves are particularly steep with less than 200 samples and still growing with more than 400 samples. In practice, a dataset including 200 samples is often much more than what is actually available.

The difficulties in tuning the hyper-parameters are problematic for the robustness of the prognostics model and the stability of the reported results. They can also have an impact on the stability of the feature

selection when sparse models are used. Avoiding these instabilities is particularly important in the context of cancer research where it could reduce the confidence of clinicians in the proposed prognostic models and the selected features.

Instabilities in prognostic models and their associated results are also problematic when we want to compare different prognostic models. The negative bias and the increased variance may not give a correct idea of the true performance of a each model. We could totally argue that the difficulties of finding the best hyper-parameters are part of the model performances. A model with very complex hyper-parameters to tune is not a good model. Models simpler to tune or with less hyper-parameters are thus better. We used this argument in chapter 7 in favor of random non-linear projections for survival analysis.

However, this argument does not hold when the performance of a prognostic model is used to assess the prognostic potential of a gene signature (see section 5.3). In such cases, we are not interested by the general performances of a model or its difficulties to be tuned. To be extreme, we just want a good black box to test the links between a signature and survival predictions. A black box that should also have stable results to produce fair comparisons between signatures.

I think that assessing the quality of the hyper-parameters selection and improving it could improve the robustness of models and the stability of their results. It can also simplify their comparison and improve the validation of gene signatures.

Some works have already been done in automatic tuning of hyper-parameters using random search [9] or Bayesian optimization [120, 131]. These methods are assessed in terms of predictive performances, but nothing as already been done to analyse the stability of the resulting models. Moreover, those results are reported on large datasets and may not apply to very small ones. In particular, I think that we should investigate the impact of using heuristics or no tuning.

11.1.2 Heterogeneity of cancer

Cancer is an heterogeneous disease with many subtypes and many factors explaining the survival of patients. For example, breast cancer is often described with three subtypes: ER+/HER2-, ER-/HER2- and HER2+. Results showed that the biological processes associated with clinical outcome depend on these subtypes [36]. Some prognostic models already take into account those subtypes adapting the prognosis for each [143, 61]. The prognosis or the responses to chemotherapy could also change with mutations such as the P53 gene mutation [83, 134].

Similar molecular subtypes classification are defined for many cancers (*e.g.*, colon cancer [89]).

Other factors could also have an impact on the survival such as the nodal status, the age, the treatment, the size of the tumor, the grade, *etc.* If these informations should be used when predicting the outcome, it can also be viewed as confounding factors problematic when assessing the performances of new prognostic models. For example in chapter 10, we decided to report results while removing the node positive and treated patients to have a more homogeneous population.

However, the best strategy to assess new prognostic models in the presence of so many potential confounding factors is unclear. Removing them to have a more homogeneous population can give better and clearer results. But, the more patients we have, the better is the assessment. Something in between could be done with a stratification of the patients: having a unique hazard ratio but with different baseline hazards depending on these factors. We choose to treat them as confounding factors in chapter 10, but we could introduce them in the models as features. We could even introduce them as an additional supervision similarly as what we have proposed in the Coxlogit model (see chapter 6).

I think that an assessment of these different strategies should be performed to know:

- what are the advantages and risks
- what is the best strategy depending on the objectives of the study
- what are the risks of overfitting while choosing the best populations
- how to take into account the fact that these factors are often a discretization of continuous factors with no clear cuts

I also think that we are too often focus on trying to prove that a new gene signature or a new model is better than others. The multiplications of these prognostic models and subtype classifications tends diminish their overall importance and the interest of these gene signatures in the domain. However, I think that these models are important with the insights they give on the tumor biology and its heterogeneity. We should focus more on the biological interpretation of the results and less on the hazard ratio of new models (especially since we developed the balanced hazard ratio). This is particularly true since none of these models will be used alone in a clinical settings.

I think that improving the medical guidelines is another task that should be tackled independently of the validation of new gene signatures.

Haibe-Kains *et al.* [61] proposed a meta model using a fuzzy classification of patients in subtypes combined with independent models for each. A perspective of a real improvement of the survival prediction could be to generalize this process automatically. We could learn meta models combining clinical factors and prognostic models using them as features representing meta-genes, prognostic pathways, *etc.* We could see prognostic models such as Gene70 (MammaPrint) [141], Gene76 [143], *etc* as a biologically relevant feature extraction or dimensionality reduction.

To have a survival prediction that depends on subtypes or clinical factors, meta models should be non-linear, such as in [143, 61]. While the benefits of multivariate non-linear survival models are yet unclear on microarray, these non-linear models such as the one presented in chapter 7 are more promising in this settings.

Chapter A

Supplementary hypoxia signature

A.1 Cell lines

ID	Cell line	Organ	Disease
1	MCF-7	Breast	Adenocarcinoma
2	MDA-MB-231	Breast	Adenocarcinoma
3	T47D	Breast	Ductal carcinoma
4	A549	Lung	Carcinoma
5	Widr	Colon	Colorectal adenocarcinoma
6	HCT116 WTP53	Colon	Colorectal carcinoma
7	HCT116 -/- P53	Colon	Colorectal carcinoma
8	HT29	Colon	Colorectal adenocarcinoma
9	Colo-205	Colon	Colorectal adenocarcinoma
10	LoVo	Colon	Colorectal adenocarcinoma
11	HCT15	Colon	Colorectal adenocarcinoma
12	SiHa	Cervix	Squamous cell carcinoma
13	PC3	Prostate	Adenocarcinoma
14	U373	Brain	Glioblastoma
15	HepG2	Liver	Hepatocellular carcinoma
16	Hep3B	Liver	Hepatocellular carcinoma
17	PLC/PRF/5	Liver	Hepatoma
18	SK-HEP-1	Liver	Adenocarcinoma
19	A498	Kidney	Carcinoma
20	HT1080	Connective tissue	Fibrosarcoma

Table A.1: List of Human Tumor Cells used for Microarray Analysis. Cancer cells were acquired from the ATCC where they are regularly authenticated. Cells were stored according to the supplier's instructions and used within 6 months after resuscitation of frozen aliquots.

A.2 Outlier detection on the cell lines data

To check for outlier in the cell lines data, we used the methodology proposed by Boulesteix *et al.* [16]. The main idea to detect outliers is to compare two rankings of the features r^* and r, respectively with and without a feature transformation. The two rankings are computed with the same feature ranking method, here a t-test. The difference comes from a feature transformation which is applied on the data before computing the ranking r^*. The objective of this transformation is to mitigate the influence of outliers. The transformation is applied after the data normalization and can be computed as follows:

$$x_{ij}^* = \left[\log\left(\frac{\phi(x_{ij})+\epsilon}{1-\phi(x_{ij})+\epsilon}\right)+\epsilon^*\right]/(2\epsilon^*) \tag{A.1}$$

where ϕ stands for the standard normal cumulative distribution function, ϵ is a parameter and $\epsilon^* = \log((1+\epsilon)/\epsilon)$. Royston and Sauerbrei [109] who proposed this transformation recommend the value $\epsilon = 0.01$ which is the one used here. The shape of the transformation can be seen on figure A.1. The function is nearly linear on the interval $[-2.8, 2.8]$ where most of the observations lie after the normalization. The effect of outliers is limited as the function is bounded between 0 and 1.

The two ranking r and r^* are thus respectively the rankings before and after this transformation. These two rankings should be very close when the data are not affected with outliers or extreme values. Figure A.2 compare the rankings of the best features ($r_j < 100$ or $r_j^* < 100$) on the cell lines data. The rankings are computed here with a t-test between the normoxia and cycling hypoxia samples. Similar results are obtained with normoxia and continuous hypoxia. Having all features close to the diagonal shows that there are almost no differences between the ranking before and after the transformation.

To have a better idea of the ranks discrepancy, Boulesteix *et al.* [16] proposed to compute $\Delta r_j = \frac{(r_j^* - r_j)}{min(r_j^*, r_j)}$ for each feature j. A high absolute value of Δr_j is obtained for the features affected by extreme values. For these features, the rankings are much better with either the transformed or the original data.

Figure A.3 shows the distribution of Δr_j for all features. Most of the Δr_j are really close to 0 showing a good concordance between the two ranking. None of the best features ($r_j < 100$ or $r_j^* < 100$) was observed with $|\Delta r_j| > 1$. Such very good Δr_j were observed by Boulesteix *et al.* [16] only in a few of the best datasets they reported. Those results confirms the very good quality of the cell lines data.

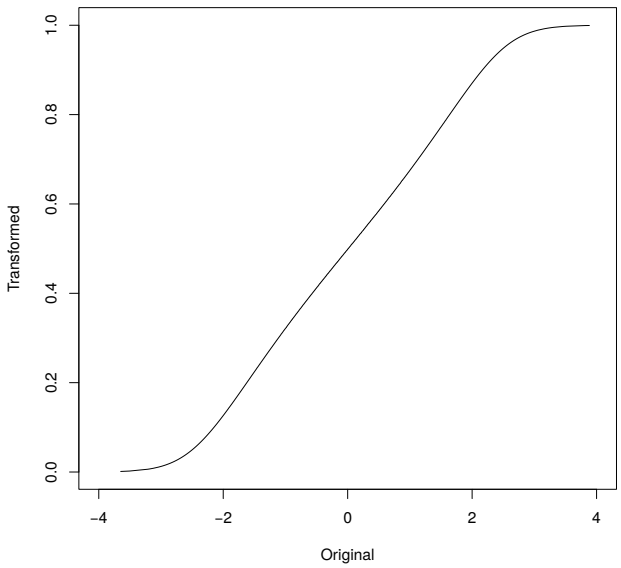


Figure A.1: Shape of the function used for data transformation.

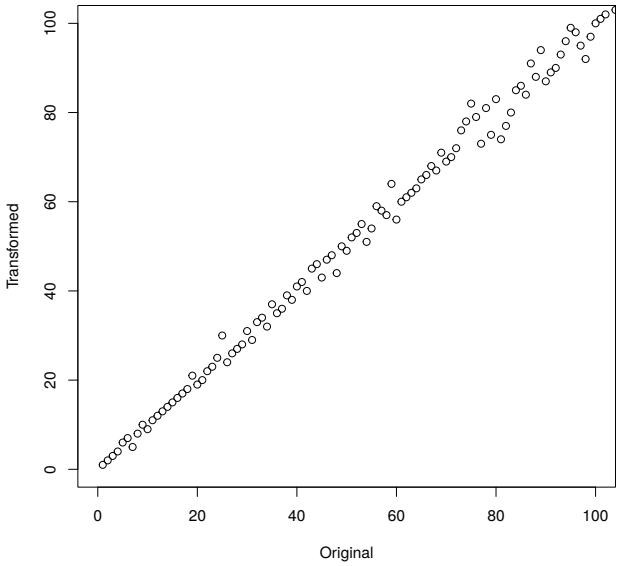


Figure A.2: Comparison between the two rankings: on the original and on the transformed features. The differences between the rankings are very small showing that the data are not affected by extreme values.

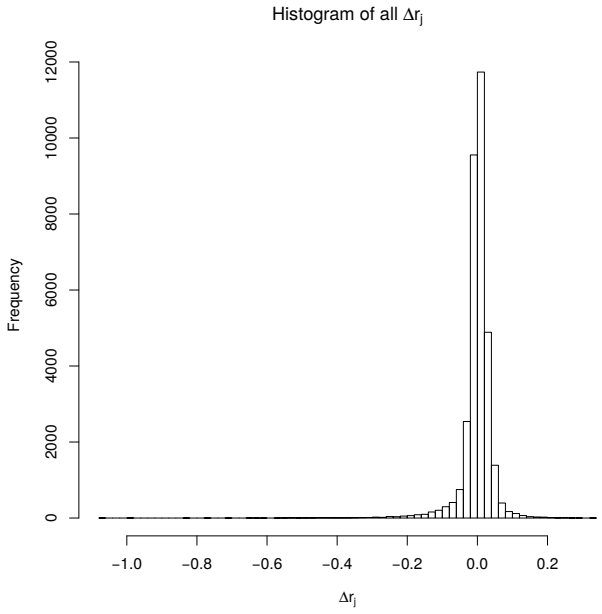


Figure A.3: Distribution of the Δr_j of all features in the data set.

A.3 CycHyp signatures

A.3.1 CycHyp signature in Affymetrix HGU1.0ST microarray platform

	Probeset	Entrez Id	GenBank	Symbol
1	**8018860**	**332**	**NM_001168**	**BIRC5**
2	**8064156**	**84619**	**NM_032527**	**ZGPAT** ∗
3	**8138912**	**23658**	**NM_012322**	**LSM5** §
4	**7921786**	**5202**	**NM_012394**	**PFDN2**
5	**8165011**	**2219**	**NM_002003**	**FCN1**
6	**7964262**	**4666**	**NM_001113201**	**NACA** ∗
7	**7949792**	**5790**	**NM_005608**	**PTPRCAP** ♯
8	**8034101**	**11018**	**NM_006858**	**TMED1**
9	**8168087**	**3476**	**NM_001551**	**IGBP1**
10	**7963575**	**1975**	**NM_001417**	**EIF4B** §
11	8124397	3006	NM_005319	HIST1H1C ♯
12	7975989	81892	NM_031210	SLIRP §
13	8127692	3351	NM_000863	HTR1B
14	8127087	2940	NM_000847	GSTA3
15	7941122	29901	NM_013299	SAC3D1
16	7998692	4913	NM_002528	NTHL1
17	8073623	758	NM_001044370	MPPED1
18	8014865	4761	NM_006160	NEUROD2 ∗
19	8005726	3768	NM_021012	KCNJ12
20	7966631	64211	NM_022363	LHX5 ∗
21	8037853	54958	NM_017854	TMEM160
22	8104136	3166	NM_018942	HMX1 ∗
23	7948606	746	NM_014206	C11orf10 ♯
24	8044773	8685	NM_006770	MARCO
25	7947015	7251	NM_006292	TSG101
26	7931553	8433	NM_003577	UTF1 ∗
27	7956876	84298	NM_032338	LLPH
28	8117372	8334	NM_003512	HIST1H2AC ♯
29	8001329	869	NM_004352	CBLN1
30	8027205	51079	NM_015965	NDUFA13
31	8042896	3196	NM_016170	TLX2 ∗
32	7911532	54998	NM_017900	AURKAIP1
33	8039923	54998	NM_017900	AURKAIP1
34	7992043	65990	BC001181	FAM173A
35	8063074	90204	NM_080603	ZSWIM1 ∗
36	7992191	23430	NM_012217	TPSD1
37	8108435	7322	NM_181838	UBE2D2
38	8165309	8721	NM_003792	EDF1 ∗
39	7946267	63875	NM_022061	MRPL17
40	7945536	51286	NM_016564	CEND1
41	8159609	8636	NM_003731	SSNA1 ♯
42	8005471	6234	NM_001031	RPS28 ♯, §
43	8025395	6234	NM_001031	RPS28
44	7942824	6234	NM_001031	RPS28
45	8170753	26576	NM_014370	SRPK3
46	8032718	1613	NM_001348	DAPK3
47	7967067	8655	NM_001037495	DYNLL1
48	8159654	25920	NM_015456	COBRA1 ∗
49	8011212	6391	NM_003001	SDHC
50	8011968	51003	NM_016060	MED31 ∗

Continued on Next Page...

Table A.2 – Continued

	Probeset	**Entrez Id**	**GenBank**	**Symbol**
51	7977440	9834	NR_026800	KIAA0125
52	8016508	11267	NM_007241	SNF8 ∗
53	8168567	5456	NM_000307	POU3F4 ∗
54	8086317	64689	NM_031899	GORASP1
55	8052834	54980	BC005079	C2orf42
56	8073334	9978	NM_014248	RBX1 ♯
57	7915846	8569	NM_003684	MKNK1
58	8071920	6634	NM_004175	SNRPD3 §
59	8032371	81926	NM_031213	FAM108A1
60	7924884	8290	NM_003493	HIST3H3
61	8006845	6143	NM_000981	RPL19 §
62	7946812	6207	NM_001017	RPS13 ♯, §
63	7949015	65998	NM_001144936	C11orf95 ∗
64	8009784	51081	NM_015971	MRPS7 §
65	8174509	2787	NM_005274	GNG5
66	7906235	5546	NM_005973	PRCC §
67	8020179	57132	NM_020412	CHMP1B
68	7947450	4005	NM_005574	LMO2
69	8064370	6939	NM_004609	TCF15 ∗
70	7955896	22818	NM_016057	COPZ1
71	8137805	8379	NM_003550	MAD1L1 ♯
72	8117334	8359	NM_003538	HIST1H4A ♯
73	8117368	8364	NM_003542	HIST1H4C ♯
74	7977507	85495	NR_002312	RPPH1 §
75	7949410	378938	BC018448	MALAT1
76	8150433	157848	NM_152568	NKX6-3 ∗
77	8071168	29797	NR_024583	POM121L8P
78	7989611	84191	NM_032231	FAM96A
79	7980859		NM_001080113	
80	8032782	126259	NM_144615	TMIGD2
81	8110861	64979	NM_032479	MRPL36 §
82	7901687	199964	NM_182532	TMEM61
83	7916130	112970	NM_138417	KTI12
84	8048712	440934	BC033986	LOC440934
85	8018993	146713	NM_001082575	RBFOX3 §
86	8032601	84839	NM_032753	RAX2
87	8010719	201255	NM_144999	LRRC45
88	8036584	3963	NM_002307	LGALS7
89	8133209	441251	NR_003666	SPDYE7P
90	8159501	286256	NM_178536	LCN12
91	8028546	3963	NM_002307	LGALS7
92	8065013		ENST00000427835	
93	8018502	201292	NM_173547	TRIM65 ∗
94	7903294	64645	NM_033055	HIAT1
95	7989473	388125	NM_001007595	C2CD4B
96	8054449	644903	AK095987	FLJ38668
97	8081867	51300	NM_016589	TIMMDC1
98	7934544	118881	NM_144589	COMTD1
99	7968260	219409	NM_145657	GSX1 ∗
100	8022952	56853	NM_020180	CELF4 §

Table A.2: CycHyp gene list in HGU1.0st

A.3.2 CycHyp signature in Affymetrix HGU133a microarray platform

	Probe	Symbol	Gene.title
1	**202095_s_at**	**BIRC5**	**baculoviral IAP repeat-containing 5**
2	**221848_at**	**ZGPAT**	**zinc finger, CCCH-type with G patch domain**
3	**202903_at**	**LSM5**	**LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae)**
4	**218336_at**	**PFDN2**	**prefoldin subunit 2**
5	**205237_at**	**FCN1**	**ficolin (collagen/fibrinogen domain containing) 1**
6	**200735_x_at**	**NACA**	**nascent polypeptide-associated complex alpha subunit**
7	**204960_at**	**PTPRCAP**	**protein tyrosine phosphatase, receptor type, C-associated protein**
8	**203679_at**	**TMED1**	**transmembrane emp24 protein transport domain containing 1**
9	**202105_at**	**IGBP1**	**immunoglobulin (CD79A) binding protein 1**
10	**211938_at**	**EIF4B**	**eukaryotic translation initiation factor 4B**
11	208046_at	HIST2H4B / HIST4H4 / HIST2H4A / HIST1H4L / HIST1H4E / HIST1H4B / HIST1H4H / HIST1H4C / HIST1H4J / HIST1H4K / HIST1H4F / HIST1H4D / HIST1H4A / HIST1H4I	histone cluster 2, H4b / histone cluster 4, H4 / histone cluster 2, H4a / histone cluster 1, H4l / histone cluster 1, H4e / histone cluster 1, H4b / histone cluster 1, H4h / histone cluster 1, H4c / histone cluster 1, H4j / histone cluster 1, H4k / histone cluster 1, H4f / histone cluster 1, H4d / histone cluster 1, H4a / histone cluster 1, H4i
12	202094_at	BIRC5	baculoviral IAP repeat-containing 5
13	209398_at	HIST1H1C	histone cluster 1, H1c
14	205967_at	HIST2H4B / HIST4H4 / HIST2H4A / HIST1H4L / HIST1H4E / HIST1H4B / HIST1H4H / HIST1H4C / HIST1H4J / HIST1H4K / HIST1H4F / HIST1H4D / HIST1H4A / HIST1H4I	histone cluster 2, H4b / histone cluster 4, H4 / histone cluster 2, H4a / histone cluster 1, H4l / histone cluster 1, H4e / histone cluster 1, H4b / histone cluster 1, H4h / histone cluster 1, H4c / histone cluster 1, H4j / histone cluster 1, H4k / histone cluster 1, H4f / histone cluster 1, H4d / histone cluster 1, H4a / histone cluster 1, H4i
15	221434_s_at	C14orf156	chromosome 14 open reading frame 156
16	210799_at	HTR1B	5-hydroxytryptamine (serotonin) receptor 1B
17	222102_at	GSTA3	glutathione S-transferase alpha 3
18	205449_at	SAC3D1	SAC3 domain containing 1
19	210334_x_at	BIRC5	baculoviral IAP repeat-containing 5
20	209731_at	NTHL1	nth endonuclease III-like 1 (E. coli)
21	206436_at	MPPED1	metallophosphoesterase domain containing 1
22	210271_at	NEUROD2	neurogenic differentiation 2
23	207110_at	KCNJ12	potassium inwardly-rectifying channel, subfamily J, member 12
24	208333_at	LHX5	LIM homeobox 5
25	202904_s_at	LSM5	LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae)
26	219219_at	TMEM160	transmembrane protein 160
27	207353_s_at	HMX1	H6 family homeobox 1
28	214386_at	HMX1	H6 family homeobox 1
29	218213_s_at	C11orf10	chromosome 11 open reading frame 10
30	205819_at	MARCO	macrophage receptor with collagenous structure
31	201758_at	TSG101	tumor susceptibility gene 101
32	208275_x_at	UTF1	undifferentiated embryonic cell transcription factor 1
33	211747_s_at	LSM5	LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae)
34	209779_at	LLPH	LLP homolog, long-term synaptic facilitation (Aplysia)
35	215071_s_at	HIST1H2AC	histone cluster 1, H2ac
36	205747_at	CBLN1	cerebellin 1 precursor
37	220864_s_at	NDUFA13	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 13

Continued on Next Page...

Table A.3 – Continued

	Probe	Symbol	Gene.title
38	207410_s_at	TLX2	T-cell leukemia homeobox 2
39	218580_x_at	AURKAIP1	aurora kinase A interacting protein 1
40	219709_x_at	FAM173A	family with sequence similarity 173, member A
41	217592_at	ZSWIM1	zinc finger, SWIM-type containing 1
42	214568_at	TPSD1	tryptase delta 1
43	201344_at	UBE2D2	ubiquitin-conjugating enzyme E2D 2 (UBC4/5 homolog, yeast)
44	209058_at	EDF1	endothelial differentiation-related factor 1
45	222216_s_at	MRPL17	mitochondrial ribosomal protein L17
46	219591_at	CEND1	cell cycle exit and neuronal differentiation 1
47	210378_s_at	SSNA1	Sjogren syndrome nuclear autoantigen 1
48	208904_s_at	RPS28	ribosomal protein S28
49	206216_at	SRPK3	SRSF protein kinase 3
50	203891_s_at	DAPK3	death-associated protein kinase 3
51	200703_at	DYNLL1	dynein, light chain, LC8-type 1
52	202757_at	COBRA1	cofactor of BRCA1
53	210131_x_at	SDHC	succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa
54	219318_x_at	MED31	mediator complex subunit 31
55	206478_at	KIAA0125	KIAA0125
56	218391_at	SNF8	SNF8, ESCRT-II complex subunit, homolog (S. cerevisiae)
57	207694_at	POU3F4	POU class 3 homeobox 4
58	201345_s_at	UBE2D2	ubiquitin-conjugating enzyme E2D 2 (UBC4/5 homolog, yeast)
59	211049_at	TLX2	T-cell leukemia homeobox 2
60	208903_at	RPS28	ribosomal protein S28
61	215749_s_at	GORASP1	golgi reassembly stacking protein 1, 65kDa
62	219128_at	C2orf42	chromosome 2 open reading frame 42
63	218117_at	RBX1	ring-box 1
64	209467_s_at	MKNK1	MAP kinase interacting serine/threonine kinase 1
65	202004_x_at	SDHC	succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa
66	208902_s_at	RPS28	ribosomal protein S28
67	202567_at	SNRPD3	small nuclear ribonucleoprotein D3 polypeptide 18kDa
68	221267_s_at	FAM108A1	family with sequence similarity 108, member A1
69	208572_at	HIST3H3	histone cluster 3, H3
70	200029_at	RPL19	ribosomal protein L19
71	200018_at	RPS13	ribosomal protein S13
72	218641_at	C11orf95	chromosome 11 open reading frame 95
73	201343_at	UBE2D2	ubiquitin-conjugating enzyme E2D 2 (UBC4/5 homolog, yeast)
74	208635_x_at	NACA	nascent polypeptide-associated complex alpha subunit
75	217932_at	MRPS7	mitochondrial ribosomal protein S7
76	207157_s_at	GNG5	guanine nucleotide binding protein (G protein), gamma 5
77	208938_at	PRCC	papillary renal cell carcinoma (translocation-associated)
78	218178_s_at	CHMP1B	chromatin modifying protein 1B
79	204249_s_at	LMO2	LIM domain only 2 (rhombotin-like 1)
80	211937_at	EIF4B	eukaryotic translation initiation factor 4B
81	207306_at	TCF15	transcription factor 15 (basic helix-loop-helix)
82	217726_at	COPZ1	coatomer protein complex, subunit zeta 1
83	204857_at	MAD1L1	MAD1 mitotic arrest deficient-like 1 (yeast)
84	218177_at	CHMP1B	chromatin modifying protein 1B
85	209059_s_at	EDF1	endothelial differentiation-related factor 1
86	219599_at	EIF4B	eukaryotic translation initiation factor 4B
87	203890_s_at	DAPK3	death-associated protein kinase 3

Table A.3: CycHyp signature in Affymetrix HGU133a microarray platform

A.4 ContHyp signatures

A.4.1 ContHyp signature in Affymetrix HGU1.0ST microarray platform

	Probeset	Entrez Id	GenBank	Symbol
1	7948606	746	NM_014206	C11orf10
2	8043283	55818	NM_018433	KDM3A
3	8025395	6234	NM_001031	RPS28
4	8139706	23480	NM_014302	SEC61G
5	7942824	6234	NM_001031	RPS28
6	8005471	6234	NM_001031	RPS28
7	8048489	55139	NM_018089	ANKZF1
8	7994737	226	NM_000034	ALDOA
9	7934278	5033	NM_000917	P4HA1
10	8102518	401152	NM_001170330	C4orf3
11	8117334	8359	NM_003538	HIST1H4A
12	8074969	1652	NM_001355	DDT
13	8044766	51141	NM_016133	INSIG2
14	7937476	6181	NM_001004	RPLP2
15	8086961	5210	NM_004567	PFKFB4
16	8145454	665	NM_004331	BNIP3L
17	8113981	8974	NM_004199	P4HA2
18	8162142	81689	NM_030940	ISCA1
19	8007992	3837	NM_002265	KPNB1
20	7928308	54541	NM_019058	DDIT4
21	8073334	9978	NM_014248	RBX1
22	8124397	3006	NM_005319	HIST1H1C
23	8153459	65263	NM_023078	PYCRL
24	7916568		AF263547	
25	7955117	23519	NM_012404	ANP32D
26	8098604	353322	NM_181726	ANKRD37
27	8121076	10957	NM_006813	PNRC1
28	7921076	54865	NM_182679	GPATCH4
29	7908879	8497	NM_015053	PPFIA4
30	8103518	23520	NM_012403	ANP32C
31	8050591	91942	NM_174889	NDUFAF2
32	8172154	6187	NM_002952	RPS2
33	7984846	1198	NM_001130028	CLK3
34	7946812	6207	NM_001017	RPS13
35	7982531	8125	NM_006305	ANP32A
36	8119898	7422	NM_001025366	VEGFA
37	8004331	9744	NM_014716	ACAP1
38	8159441	29085	NM_001135861	PHPT1
39	8168500	5230	NM_000291	PGK1
40	7938890	10196	NM_005788	PRMT3
41	7930398	4601	NM_005962	MXI1
42	7997740	81631	NM_022818	MAP1LC3B
43	8004360	147040	NM_001002914	KCTD11
44	7909782	51018	NM_016052	RRP15
45	7949792	5790	NM_005608	PTPRCAP
46	8124385	8366	NM_003544	HIST1H4B
47	8117368	8364	NM_003542	HIST1H4C
48	8081241	84319	NM_032359	C3orf26
49	8050079	246243	NM_002936	RNASEH1
50	8005765	26118	NM_015626	WSB1
51	7924491	64853	NM_022831	AIDA
52	8133273		ENST00000455206	
53	8124391	8335	NM_003513	HIST1H2AB
54	8159609	8636	NM_003731	SSNA1
55	7957890	27340	NM_014503	UTP20
56	7933582	100287932	NM_006327	TIMM23
57	8153002	10397	NM_001135242	NDRG1
58	7926037	5209	NM_004566	PFKFB3

Continued on Next Page...

Table A.4 – Continued

	Probeset	Entrez Id	GenBank	Symbol
59	8082066	26355	NM_014367	FAM162A
60	8042962	9801	NM_014763	MRPL19
61	8090678	11222	NM_007208	MRPL3
62	7977507	85495	NR_002312	RPPH1
63	8007397	10197	NM_176863	PSME3
64	7998902	54985	NM_017885	HCFC1R1
65	8117372	8334	NM_003512	HIST1H2AC
66	7997230	5713	NM_002811	PSMD7
67	7915485	10969	NM_006824	EBNA1BP2
68	8113873	3094	NM_005340	HINT1
69	7958152	5223	NM_002629	PGAM1
70	7947867	5702	NM_002804	PSMC3
71	7964460	1649	NM_004083	DDIT3
72	7928395	170384	NM_173540	FUT11
73	8163629	944	NM_001244	TNFSF8
74	7965486	51134	NM_016122	CCDC41
75	8136179	23008	AF277175	KLHDC10
76	8095870	901	NM_004354	CCNG2
77	8127526	6170	NM_001000	RPL39
78	8174710	6170	NM_001000	RPL39
79	8137517	3361	NM_024012	HTR5A
80	7929624	5223	NM_002629	PGAM1
81	8052331	87178	NM_033109	PNPT1
82	8015969	7343	NM_014233	UBTF
83	8069168	386685	NM_198699	KRTAP10-12
84	7941087	5526	NM_006244	PPP2R5B
85	8026875	26780	NR_000012	SNORA68
86	8027621	2821	NM_000175	GPI
87	8130539	117289	NM_054114	TAGAP
88	8004691	92162	NM_203411	TMEM88
89	7962183	205	NM_001005353	AK4
90	8137805	8379	NM_003550	MAD1L1
91	8124388	8358	NM_003537	HIST1H3B
92	8083223	205428	NM_173552	C3orf58
93	8113305	1105	NM_001270	CHD1
94	8169659	4694	NM_004541	NDUFA1
95	8046408	5163	NM_002610	PDK1
96	8053599	23559	NM_012477	WBP1
97	8043377	23559	NM_012477	WBP1
98	7960878	642559	GU480887	POU5F1P3
99	7959023	643246	NM_001085481	MAP1LC3B2
100	8073148	468	NM_001675	ATF4

Table A.4: ContHyp gene list in HGU1.0st

A.4.2 ContHyp signature in Affymetrix HGU133a microarray platform

	Probe	Symbol	Gene.title
1	208576_s_at	HIST1H3F / HIST1H3B / HIST1H3H / HIST1H3J / HIST1H3G / HIST1H3I / HIST1H3E / HIST1H3C / HIST1H3D / HIST1H3A / HIST1H2AD	histone cluster 1, H3f / histone cluster 1, H3b / histone cluster 1, H3h / histone cluster 1, H3j / histone cluster 1, H3g / histone cluster 1, H3i / histone cluster 1, H3e / histone cluster 1, H3c / histone cluster 1, H3d / histone cluster 1, H3a / histone cluster 1, H2ad
2	218944_at	PYCRL	pyrroline-5-carboxylate reductase-like
3	202464_s_at	PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3
4	214978_s_at	PPFIA4	protein tyrosine phosphatase, receptor type, f polypeptide (PTPRF), interacting protein (liprin), alpha 4
5	219037_at	RRP15	ribosomal RNA processing 15 homolog (S. cerevisiae)
6	217670_at	RPLP2	ribosomal protein, large, P2
7	220596_at	GPATCH4	G patch domain containing 4
8	200738_s_at	PGK1	phosphoglycerate kinase 1
9	213507_s_at	KPNB1	karyopherin (importin) beta 1
10	200886_s_at	PGAM1	phosphoglycerate mutase 1 (brain)
11	201705_at	PSMD7	proteasome (prosome, macropain) 26S subunit, non-ATPase, 7
12	210512_s_at	VEGFA	vascular endothelial growth factor A
13	214764_at	RRP15	ribosomal RNA processing 15 homolog (S. cerevisiae)
14	204258_at	CHD1	chromodomain helicase DNA binding protein 1
15	203465_at	MRPL19	mitochondrial ribosomal protein L19
16	202733_at	P4HA2	prolyl 4-hydroxylase, alpha polypeptide II
17	218497_s_at	RNASEH1	ribonuclease H1
18	207543_s_at	P4HA1	prolyl 4-hydroxylase, alpha polypeptide I
19	209273_s_at	ISCA1	iron-sulfur cluster assembly 1 homolog (S. cerevisiae)
20	200632_s_at	NDRG1	N-myc downstream regulated 1
21	635_s_at	PPP2R5B	protein phosphatase 2, regulatory subunit B', beta
22	209725_at	UTP20	UTP20, small subunit (SSU) processome component, homolog (yeast)
23	208903_at	RPS28	ribosomal protein S28
24	218274_s_at	ANKZF1	ankyrin repeat and zinc finger domain containing 1
25	202692_s_at	UBTF	upstream binding transcription factor, RNA polymerase I
26	218117_at	RBX1	ring-box 1
27	204611_s_at	PPP2R5B	protein phosphatase 2, regulatory subunit B', beta
28	210111_s_at	KLHDC10	kelch domain containing 10
29	200018_at	RPS13	ribosomal protein S13
30	204960_at	PTPRCAP	protein tyrosine phosphatase, receptor type, C-associated protein
31	200779_at	ATF4	activating transcription factor 4 (tax-responsive enhancer element B67)
32	210265_x_at	POU5F1P3	POU class 5 homeobox 1 pseudogene 3
33	220942_x_at	FAM162A	family with sequence similarity 162, member A
34	213406_at	WSB1	WD repeat and SOCS box-containing 1
35	202364_at	MXI1	MAX interactor 1
36	201051_at	ANP32A	acidic (leucine-rich) nuclear phosphoprotein 32 family, member A
37	200909_s_at	RPLP2	ribosomal protein, large, P2
38	204857_at	MAD1L1	MAD1 mitotic arrest deficient-like 1 (yeast)
39	214881_s_at	UBTF	upstream binding transcription factor, RNA polymerase I
40	221478_at	BNIP3L	BCL2/adenovirus E1B 19kDa interacting protein 3-like
41	221479_s_at	BNIP3L	BCL2/adenovirus E1B 19kDa interacting protein 3-like
42	211527_x_at	VEGFA	vascular endothelial growth factor A
43	209034_at	PNRC1	proline-rich nuclear receptor coactivator 1
44	201295_s_at	WSB1	WD repeat and SOCS box-containing 1

Continued on Next Page...

Table A.5 – Continued

	Probe	Symbol		Gene.title
45	208046_at	HIST2H4B	/	histone cluster 2, H4b / histone cluster 4, H4 / histone cluster 2, H4a / histone cluster 1, H4l / histone cluster 1, H4e / histone cluster 1, H4b / histone cluster 1, H4h / histone cluster 1, H4c / histone cluster 1, H4j / histone cluster 1, H4k / histone cluster 1, H4f / histone cluster 1, H4d / histone cluster 1, H4a / histone cluster 1, H4i
		HIST4H4	/	
		HIST2H4A	/	
		HIST1H4L	/	
		HIST1H4E	/	
		HIST1H4B	/	
		HIST1H4H	/	
		HIST1H4C	/	
		HIST1H4J	/	
		HIST1H4K	/	
		HIST1H4F	/	
		HIST1H4D	/	
		HIST1H4A	/	
		HIST1H4I		
46	209398_at	HIST1H1C		histone cluster 1, H1c
47	205967_at	HIST2H4B	/	histone cluster 2, H4b / histone cluster 4, H4 / histone cluster 2, H4a / histone cluster 1, H4l / histone cluster 1, H4e / histone cluster 1, H4b / histone cluster 1, H4h / histone cluster 1, H4c / histone cluster 1, H4j / histone cluster 1, H4k / histone cluster 1, H4f / histone cluster 1, H4d / histone cluster 1, H4a / histone cluster 1, H4i
		HIST4H4	/	
		HIST2H4A	/	
		HIST1H4L	/	
		HIST1H4E	/	
		HIST1H4B	/	
		HIST1H4H	/	
		HIST1H4C	/	
		HIST1H4J	/	
		HIST1H4K	/	
		HIST1H4F	/	
		HIST1H4D	/	
		HIST1H4A	/	
		HIST1H4I		
48	221733_s_at	GPATCH4		G patch domain containing 4
49	214516_at	HIST2H4B	/	histone cluster 2, H4b / histone cluster 4, H4 / histone cluster 2, H4a / histone cluster 1, H4l / histone cluster 1, H4e / histone cluster 1, H4b / histone cluster 1, H4h / histone cluster 1, H4c / histone cluster 1, H4j / histone cluster 1, H4k / histone cluster 1, H4f / histone cluster 1, H4d / histone cluster 1, H4a / histone cluster 1, H4i
		HIST4H4	/	
		HIST2H4A	/	
		HIST1H4L	/	
		HIST1H4E	/	
		HIST1H4B	/	
		HIST1H4H	/	
		HIST1H4C	/	
		HIST1H4J	/	
		HIST1H4K	/	
		HIST1H4F	/	
		HIST1H4D	/	
		HIST1H4A	/	
		HIST1H4I		
50	200908_s_at	RPLP2		ribosomal protein, large, P2
51	217027_x_at	KPNB1		karyopherin (importin) beta 1
52	203484_at	SEC61G		Sec61 gamma subunit
53	219644_at	CCDC41		coiled-coil domain containing 41
54	218213_s_at	C11orf10		chromosome 11 open reading frame 10
55	217383_at	PGK1		phosphoglycerate kinase 1
56	220611_at	DAB1		disabled homolog 1 (Drosophila)
57	215071_s_at	HIST1H2AC		histone cluster 1, H2ac
58	200737_at	PGK1		phosphoglycerate kinase 1
59	213320_at	PRMT3		protein arginine methyltransferase 3
60	202887_s_at	DDIT4		DNA-damage-inducible transcript 4
61	210561_s_at	WSB1		WD repeat and SOCS box-containing 1
62	208826_x_at	HINT1		histidine triad nucleotide binding protein 1
63	209852_x_at	PSME3		proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki)
64	200988_s_at	PSME3		proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki)
65	201296_s_at	WSB1		WD repeat and SOCS box-containing 1
66	221798_x_at	RPS2		ribosomal protein S2
67	212171_x_at	VEGFA		vascular endothelial growth factor A
68	210378_s_at	SSNA1		Sjogren syndrome nuclear autoantigen 1
69	202298_at	NDUFA1		NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1, 7.5kDa
70	201038_s_at	ANP32A		acidic (leucine-rich) nuclear phosphoprotein 32 family, member A
71	206246_at	PFKFB4		6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4
72	203107_x_at	RPS2		ribosomal protein S2
73	208904_s_at	RPS28		ribosomal protein S28
74	208974_x_at	KPNB1		karyopherin (importin) beta 1
75	200966_x_at	ALDOA		aldolase A, fructose-bisphosphate
76	214687_x_at	ALDOA		aldolase A, fructose-bisphosphate
77	202929_s_at	DDT		D-dopachrome tautomerase
78	212433_x_at	RPS2		ribosomal protein S2
79	208695_s_at	RPL39		ribosomal protein L39

Continued on Next Page...

Table A.5 – Continued

	Probe	Symbol	Gene.title
80	217466_x_at	RPS2	ribosomal protein S2
81	202140_s_at	CLK3	CDC-like kinase 3
82	208787_at	MRPL3	mitochondrial ribosomal protein L3
83	208569_at	HIST1H2AB	histone cluster 1, H2ab
84	201323_at	EBNA1BP2	EBNA1 binding protein 2
85	213574_s_at	KPNB1	karyopherin (importin) beta 1
86	202769_at	CCNG2	cyclin G2
87	218496_at	RNASEH1	ribonuclease H1
88	45714_at	HCFC1R1	host cell factor C1 regulator 1 (XPO1 dependent)
89	207216_at	TNFSF8	tumor necrosis factor (ligand) superfamily, member 8
90	200093_s_at	HINT1	histidine triad nucleotide binding protein 1
91	212689_s_at	KDM3A	lysine (K)-specific demethylase 3A
92	208902_s_at	RPS28	ribosomal protein S28
93	209254_at	KLHDC10	kelch domain containing 10
94	221533_at	FAM162A	family with sequence similarity 162, member A
95	218118_s_at	LOC10431	translocase of inner mitochondrial membrane 23 homolog (yeast)-like
96	208975_s_at	KPNB1	karyopherin (importin) beta 1
97	221362_at	HTR5A	5-hydroxytryptamine (serotonin) receptor 5A
98	208308_s_at	GPI	glucose-6-phosphate isomerase
99	205213_at	ACAP1	ArfGAP with coiled-coil, ankyrin repeat and PH domains 1
100	209274_s_at	ISCA1	iron-sulfur cluster assembly 1 homolog (S. cerevisiae)
101	204348_s_at	AK3L2 / AK4	adenylate kinase 3-like 2 (pseudogene) / adenylate kinase 4
102	207721_x_at	HINT1	histidine triad nucleotide binding protein 1
103	213573_at	KPNB1	karyopherin (importin) beta 1
104	208538_at	ANP32C	acidic (leucine-rich) nuclear phosphoprotein 32 family, member C
105	221425_s_at	ISCA1	iron-sulfur cluster assembly 1 homolog (S. cerevisiae)
106	209566_at	INSIG2	insulin induced gene 2
107	209853_s_at	PSME3	proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki)
108	204347_at	AK3L2 / AK4	adenylate kinase 3-like 2 (pseudogene) / adenylate kinase 4
109	200987_x_at	PSME3	proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki)
110	210513_s_at	VEGFA	vascular endothelial growth factor A
111	217356_s_at	PGK1	phosphoglycerate kinase 1
112	209255_at	KLHDC10	kelch domain containing 10
113	208786_s_at	MAP1LC3B	microtubule-associated protein 1 light chain 3 beta
114	213803_at	KPNB1	karyopherin (importin) beta 1
115	201294_s_at	WSB1	WD repeat and SOCS box-containing 1
116	205212_s_at	ACAP1	ArfGAP with coiled-coil, ankyrin repeat and PH domains 1
117	206686_at	PDK1	pyruvate dehydrogenase kinase, isozyme 1
118	201267_s_at	PSMC3	proteasome (prosome, macropain) 26S subunit, ATPase, 3
119	220199_s_at	AIDA	axin interactor, dorsalization associated
120	218537_at	HCFC1R1	host cell factor C1 regulator 1 (XPO1 dependent)
121	202770_s_at	CCNG2	cyclin G2
122	201043_s_at	ANP32A	acidic (leucine-rich) nuclear phosphoprotein 32 family, member A
123	211559_s_at	CCNG2	cyclin G2

Table A.5: ContHyp signature in Affymetrix HGU133a microarray platform

A.5 Heatmap depicting the transcripts from the ContHyp signature

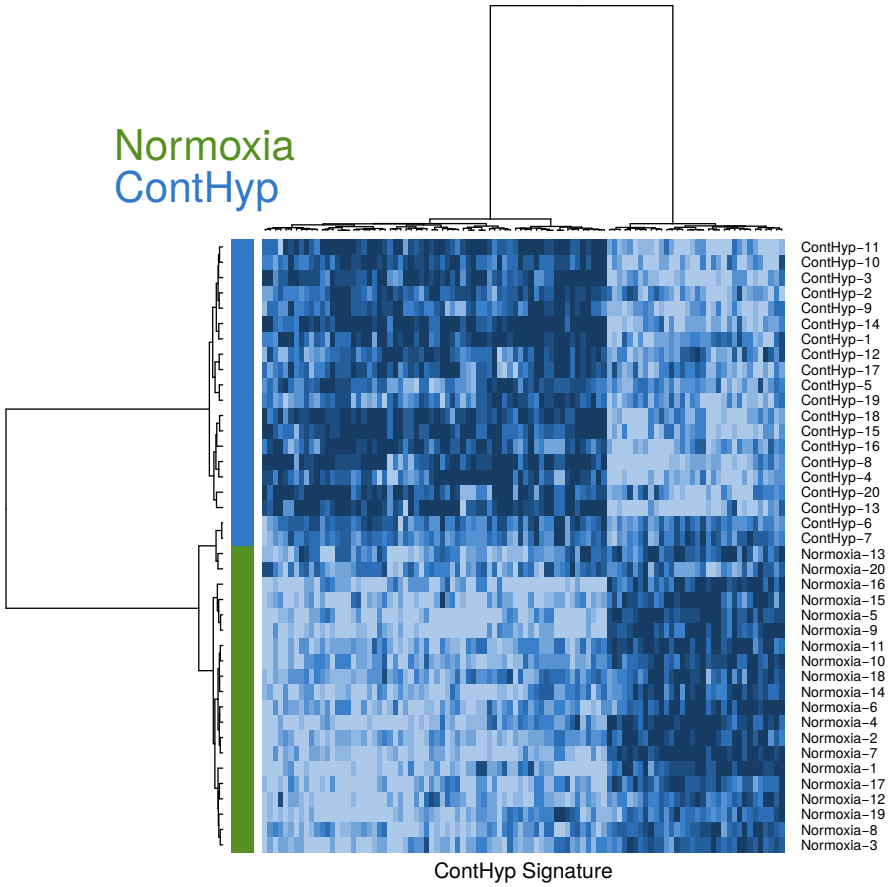


Figure A.4: Heatmap depicting the transcripts from the ContHyp signature either underexpressed (light blue) or overexpressed (dark blue) (centered to median values). Each column corresponds to a specific human Gene 1.0 ST probeset ; each line represents a specific cell line either maintained under normoxia (green) or exposed to continuous hypoxia (blue). The numbers (from 1 to 20) can be used to access information about specific cell lines in table A.1 in appendix.

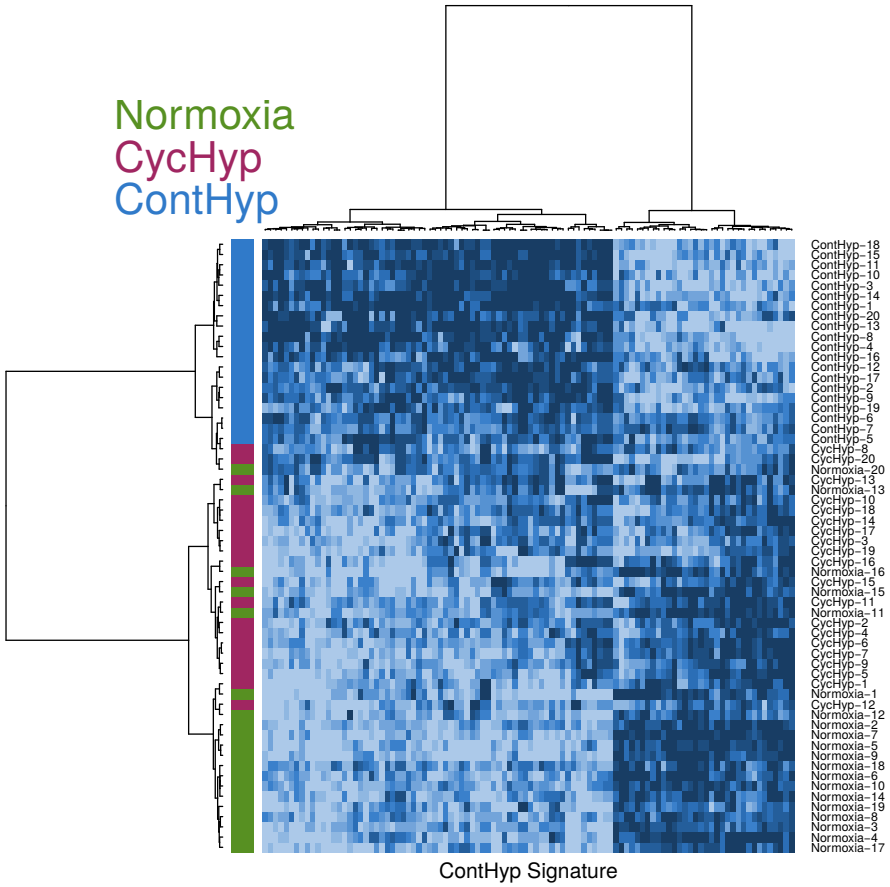


Figure A.5: Heatmap depicting the transcripts from the ContHyp signature either underexpressed (light blue) or overexpressed (dark blue) (centered to median values). Each column corresponds to a specific human Gene 1.0 ST probeset ; each line represents a specific cell line either maintained under normoxia (green), exposed to continuous hypoxia (blue) or exposed to cycling hypoxia (purple). The numbers (from 1 to 20) can be used to access information about specific cell line in table A.1 in appendix.

A.6 GSEA analysis on the CycHyp and ContHyp data

Gene sets	Size	ES	p-val	FDR q-val	FWER p-val
JISON SICKLE CELL DISEASE DN	26	0.442	0	***0.011***	0.012
ZWANG TRANSIENTLY UP BY 2ND EGF PULSE ONLY	226	0.158	0	***0.016***	0.036
REACTOME TRANSLATION	40	0.275	0.006	0.44	0.771
PECE MAMMARY STEM CELL DN	29	0.303	0.008	0.501	0.894
KRIGE RESPONSE TO TOSEDOSTAT 24HR DN	94	0.12	0.155	0.978	1
BURTON ADIPOGENESIS 5	15	0.282	0.168	0.986	1
HAMAI APOPTOSIS VIA TRAIL DN	17	0.269	0.151	0.99	1
HSIAO HOUSEKEEPING GENES	65	0.143	0.14	0.99	1
MULLIGHAN MLL SIGNATURE 2 DN	15	0.277	0.175	0.992	1
REACTOME RNA POL I TRANSCRIPTION	17	0.267	0.16	0.992	1
HUTTMANN B CLL POOR SURVIVAL UP	20	0.26	0.12	0.993	1
REACTOME PEPTIDE CHAIN ELONGATION	27	0.202	0.206	0.994	1
WANG TUMOR INVASIVENESS UP	46	0.182	0.094	0.999	1
BOUDOUKHA BOUND BY IGF2BP2	15	0.371	0.024	1	0.998
BURTON ADIPOGENESIS 6	16	0.304	0.094	1	1
LI INDUCED T TO NATURAL KILLER UP	15	0.301	0.111	1	1
KEGG UBIQUITIN MEDIATED PROTEOLYSIS	22	0.29	0.045	1	1
KEGG SYSTEMIC LUPUS ERYTHEMATOSUS	16	0.289	0.118	1	1
GOLDRATH ANTIGEN RESPONSE	19	0.286	0.078	1	1
WIERENGA STAT5A TARGETS DN	18	0.285	0.082	1	1
YAO TEMPORAL RESPONSE TO PROGESTERONE CLUSTER 17	21	0.281	0.064	1	1
LINDGREN BLADDER CANCER CLUSTER 3 UP	16	0.271	0.163	1	1
MALONEY RESPONSE TO 17AAG DN	18	0.264	0.144	1	1
UDAYAKUMAR MED1 TARGETS UP	19	0.26	0.145	1	1
AFFAR YY1 TARGETS UP	20	0.259	0.137	1	1
KEGG PATHWAYS IN CANCER	28	0.252	0.054	1	1
SWEET LUNG CANCER KRAS DN	26	0.25	0.076	1	1
REACTOME 3 UTR MEDIATED TRANSLATIONAL REGULATION	32	0.249	0.033	1	1
YAGI AML WITH T 8 21 TRANSLOCATION	20	0.249	0.147	1	1
KEGG WNT SIGNALING PATHWAY	19	0.239	0.208	1	1
GARY CD5 TARGETS UP	31	0.228	0.073	1	1
REACTOME SRP DEPENDENT COTRANSLATIONAL PROTEIN TARGETING TO MEMBRANE	33	0.227	0.064	1	1
NIKOLSKY BREAST CANCER 16P13 AMPLICON	26	0.227	0.118	1	1
MOHANKUMAR TLX1 TARGETS UP	27	0.224	0.122	1	1
NIKOLSKY BREAST CANCER 17Q21 Q25 AMPLICON	36	0.217	0.068	1	1
WONG EMBRYONIC STEM CELL CORE	46	0.196	0.062	1	1
YAGI AML WITH INV 16 TRANSLOCATION	34	0.186	0.197	1	1
MIKKELSEN ES ICP WITH H3K4ME3	54	0.175	0.076	1	1
KRIGE RESPONSE TO TOSEDOSTAT 6HR UP	42	0.167	0.192	1	1
STARK PREFRONTAL CORTEX 22Q11 DELETION DN	71	0.166	0.047	1	1
DELACROIX RAR BOUND ES	45	0.164	0.178	1	1
MOOTHA MITOCHONDRIA	65	0.152	0.099	1	1
MEISSNER BRAIN HCP WITH H3K27ME3	55	0.151	0.176	1	1
MARTENS TRETINOIN RESPONSE DN	108	0.148	0.028	1	0.999
KRIGE RESPONSE TO TOSEDOSTAT 6HR DN	87	0.141	0.073	1	1
REACTOME METABOLISM OF PROTEINS	75	0.141	0.113	1	1
MIKKELSEN MEF HCP WITH H3K27ME3	89	0.137	0.085	1	1
ACEVEDO LIVER TUMOR VS NORMAL ADJACENT TISSUE UP	94	0.129	0.103	1	1
BENPORATH EED TARGETS	115	0.104	0.196	1	1
DANG BOUND BY MYC	127	0.102	0.174	1	1

Table A.6: GSEA preranked analysis on the CycHyp signature extended to 2118 differentially expressed genes between normoxia and cycling hypoxia. The table contains the signature sizes, primary and normalized enrichment scores (ES and NES, respectively), the nominal pvalues, the false discovery rates (FDR) qvalues and the familywise error rate (FWER) for the top 50 gene sets enriched in this CycHyp signature. Only the two first gene sets are significantly enriched with a FDR qvalue < 0.1.

Gene sets	Size	ES	p-val	FDR q-val	FWER p-val
FARDIN HYPOXIA 11	17	0.712	0	0	0
LEONARD HYPOXIA	21	0.524	0	0	0.002
MENSE HYPOXIA UP	39	0.477	0	0	0
ELVIDGE HYPOXIA BY DMOG UP	46	0.459	0	0	0
ELVIDGE HIF1A AND HIF2A TARGETS DN	44	0.445	0	0	0
WINTER HYPOXIA METAGENE	46	0.443	0	0	0
QI HYPOXIA	36	0.421	0	0	0
ELVIDGE HIF1A TARGETS DN	43	0.39	0	0	0
ELVIDGE HYPOXIA UP	56	0.386	0	0	0
KRIEG HYPOXIA NOT VIA KDM3A	113	0.295	0	0	0
NAKAMURA TUMOR ZONE PERIPHERAL VS CENTRAL DN	76	0.27	0	0.002	0.015
PRAMOONJAGO SOX4 TARGETS UP	15	0.551	0	0.003	0.029
WINTER HYPOXIA UP	20	0.491	0	0.003	0.024
GROSS HYPOXIA VIA HIF1A DN	23	0.432	0	0.007	0.083
GROSS HYPOXIA VIA ELK3 AND HIF1A UP	29	0.392	0	0.007	0.079
BASAKI YBX1 TARGETS DN	35	0.359	0	0.007	0.07
MARTORIATI MDM4 TARGETS NEUROEPITHELIUM UP	23	0.432	0	0.008	0.091
MARKEY RB1 ACUTE LOF UP	24	0.419	0	0.008	0.103
SCHLOSSER MYC TARGETS AND SERUM RESPONSE DN	18	0.48	0	0.009	0.117
MANALO HYPOXIA UP	35	0.35	0	0.009	0.113
HARRIS HYPOXIA	17	0.488	0.002	0.01	0.143
REACTOME METABOLISM OF RNA	60	0.264	0.001	0.012	0.177
BOYAULT LIVER CANCER SUBCLASS G3 UP	44	0.301	0.001	0.014	0.22
BLALOCK ALZHEIMERS DISEASE DN	169	0.163	0	0.014	0.208
DANG BOUND BY MYC	170	0.162	0.002	0.016	0.253
PID HIF1 TFPATHWAY	16	0.477	0.002	0.017	0.283
DANG MYC TARGETS UP	44	0.3	0.001	0.017	0.278
SHAFFER IRF4 TARGETS IN MYELOMA VS MATURE B LYMPHOCYTE	15	0.488	0	0.018	0.304
WEI MYCN TARGETS WITH E BOX	194	0.149	0.001	0.022	0.372
KRIGE RESPONSE TO TOSEDOSTAT 24HR DN	220	0.139	0.001	0.024	0.408
REACTOME INFLUENZA LIFE CYCLE	33	0.322	0.002	0.03	0.492
NUYTTEN EZH2 TARGETS UP	78	0.217	0.002	0.03	0.501
GROSS HYPOXIA VIA ELK3 DN	29	0.339	0.002	0.031	0.535
BENPORATH MYC MAX TARGETS	120	0.176	0.001	0.032	0.558
CAIRO HEPATOBLASTOMA CLASSES UP	139	0.167	0.003	0.032	0.53
PODAR RESPONSE TO ADAPHOSTIN UP	25	0.352	0.003	0.049	0.725
KIM MYC AMPLIFICATION TARGETS UP	33	0.306	0.004	0.053	0.762
YAMAZAKI TCEB3 TARGETS DN	25	0.349	0.005	0.055	0.786
RODRIGUES THYROID CARCINOMA POORLY DIFFERENTIATED UP	113	0.168	0.005	0.064	0.845
REACTOME RNA POL I PROMOTER OPENING	16	0.415	0.006	0.069	0.87
HSIAO HOUSEKEEPING GENES	72	0.208	0.004	0.071	0.88
GARY CD5 TARGETS DN	109	0.17	0.005	0.073	0.903
DODD NASOPHARYNGEAL CARCINOMA DN	233	0.122	0.002	0.074	0.9
GRAESSMANN APOPTOSIS BY SERUM DEPRIVATION DN	28	0.322	0.006	0.075	0.899
YAO TEMPORAL RESPONSE TO PROGESTERONE CLUSTER 11	31	0.301	0.006	0.078	0.924
KRIGE RESPONSE TO TOSEDOSTAT 6HR DN	188	0.131	0.007	0.083	0.941
GRAESSMANN RESPONSE TO MC AND DOXORUBICIN DN	123	0.158	0.004	0.084	0.94
LEE BMP2 TARGETS DN	214	0.125	0.008	0.084	0.946
REACTOME RNA POL I TRANSCRIPTION	20	0.363	0.006	0.085	0.951
WANG CISPLATIN RESPONSE AND XPC DN	24	0.329	0.007	0.087	0.96
LOCKWOOD AMPLIFIED IN LUNG CANCER	27	0.312	0.009	0.098	0.975
PECE MAMMARY STEM CELL DN	29	0.297	0.011	0.099	0.978

Table A.7: GSEA preranked analysis on the ContHyp signature extended to 2065 differentially expressed genes between normoxia and continuous hypoxia. The table contains the signature sizes, primary and normalized enrichment scores (ES and NES, respectively), the nominal pvalues, the false discovery rates (FDR) qvalues and the familywise error rate (FWER). 17 gene sets (bold) were previously identified as related to hypoxia.

A.7 Overlap with other hypoxia-related gene signatures

	Hypoxia Gene Set	ps in common
1	PID HIF1APATHWAY	1
2	REACTOME REGULATION OF HYPOXIA INDUCIBLE FACTOR HIF BY OXYGEN	2
3	GROSS HYPOXIA VIA ELK3 AND HIF1A DN	1
4	REACTOME OXYGEN DEPENDENT PROLINE HYDROXYLATION OF HYPOXIA INDUCIBLE FACTOR ALPHA	2
5	WINTER HYPOXIA UP	1
6	WINTER HYPOXIA DN	1
7	ELVIDGE HYPOXIA BY DMOG DN	1
8	GROSS HYPOXIA VIA ELK3 UP	2
9	GROSS HYPOXIA VIA ELK3 ONLY DN	1
10	MANALO HYPOXIA DN	1
11	JIANG HYPOXIA NORMAL	1
12	JIANG HYPOXIA CANCER	1
13	KRIEG HYPOXIA NOT VIA KDM3A	4

Table A.8: Overlap, in terms of number of common genes, between the CycHyp signature and gene sets from the MsigDB identified as being related to hypoxia or HIF.

	Hypoxia Gene Set	ps in common
1	BIOCARTA HIF PATHWAY	1
2	PID HIF2PATHWAY	2
3	PID HIF1APATHWAY	1
4	PID HIF1 TFPATHWAY	5
5	REACTOME REGULATION OF HYPOXIA INDUCIBLE FACTOR HIF BY OXYGEN	2
6	ELVIDGE HIF1A TARGETS UP	2
7	ELVIDGE HIF1A TARGETS DN	15
8	ELVIDGE HIF1A AND HIF2A TARGETS DN	16
9	GROSS HYPOXIA VIA HIF1A ONLY	1
10	GROSS HIF1A TARGETS DN	3
11	GROSS HYPOXIA VIA HIF1A DN	4
12	GROSS HYPOXIA VIA ELK3 AND HIF1A UP	14
13	RANKIN ANGIOGENIC TARGETS OF VHL HIF2A DN	1
14	SEMENZA HIF1 TARGETS	4
15	QI HYPOXIA TARGETS OF HIF1A AND FOXA2	1
16	REACTOME OXYGEN DEPENDENT PROLINE HYDROXYLATION OF HYPOXIA INDUCIBLE FACTOR ALPHA	1
17	WINTER HYPOXIA UP	11
18	ELVIDGE HYPOXIA UP	19
19	ELVIDGE HYPOXIA DN	4
20	ELVIDGE HYPOXIA BY DMOG UP	17
21	ELVIDGE HYPOXIA BY DMOG DN	2
22	WEINMANN ADAPTATION TO HYPOXIA UP	1
23	WEINMANN ADAPTATION TO HYPOXIA DN	1
24	KONDO HYPOXIA	1
25	GROSS HYPOXIA VIA ELK3 UP	4
26	GROSS HYPOXIA VIA ELK3 DN	6
27	GROSS HYPOXIA VIA ELK3 ONLY UP	1
28	MANALO HYPOXIA DN	5
29	MANALO HYPOXIA UP	10
30	MENSE HYPOXIA UP	19
31	KIM HYPOXIA	4
32	HARRIS HYPOXIA	7
33	LEONARD HYPOXIA	12
34	JIANG HYPOXIA NORMAL	9
35	JIANG HYPOXIA CANCER	2
36	JIANG AGING HYPOTHALAMUS UP	1
37	WINTER HYPOXIA METAGENE	16
38	MIZUKAMI HYPOXIA UP	1
39	QI HYPOXIA	14
40	FARDIN HYPOXIA 9	5
41	FARDIN HYPOXIA 11	14
42	WACKER HYPOXIA TARGETS OF VHL	3
43	KRIEG HYPOXIA VIA KDM3A	1
44	KRIEG HYPOXIA NOT VIA KDM3A	27

Table A.9: Overlap, in terms of number of common genes, between the ContHyp signature and gene sets from the MsigDB identified as being related to hypoxia or HIF.

		Size	CycHyp	ContHyp
Seigneuric *et al.* [114]	Early 0%	72	0	0
	Late 0%	71	1	7
	Early 2%	34	1	0
	Late 2%	32	0	3
Starmans *et al.* [126]	Cluster 1	69	0	5
	Cluster 2	246	1	20
	Cluster 3	157	0	4
	Cluster 4	95	1	1
	Cluster 5	162	0	0
	Cluster 6	14	0	0
	Cluster 7	28	1	0
	Upregulated	780	2	32
	Downregulated	656	6	6

Table A.10: Overlap, in terms of number of common genes, between the CycHyp or ContHyp signatures and the conventional hypoxiarelated signatures described by Seigneuric *et al.* [114] and Starmans *et al.* [126].

A.8 Hypoxia signatures on different breast cancer subpopulations

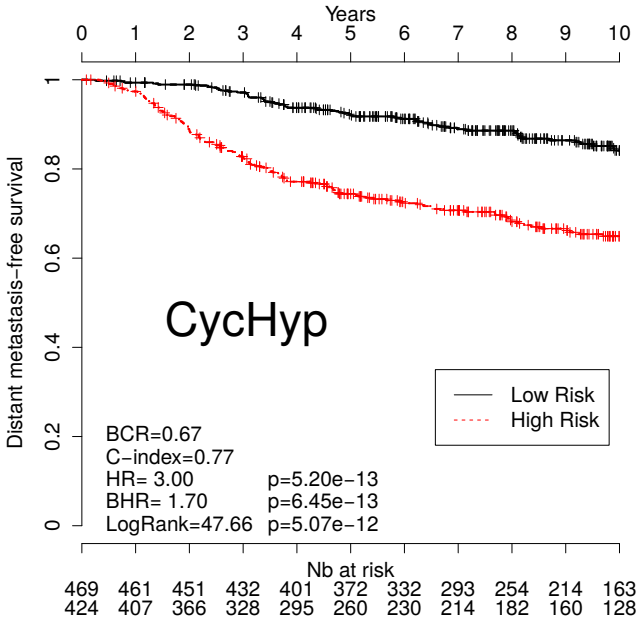


Figure A.6: Kaplan-Meier survival curves of patients with primary breast cancer, as determined by using the CycHyp signature. The validation set is reduced here to node-negative patients.

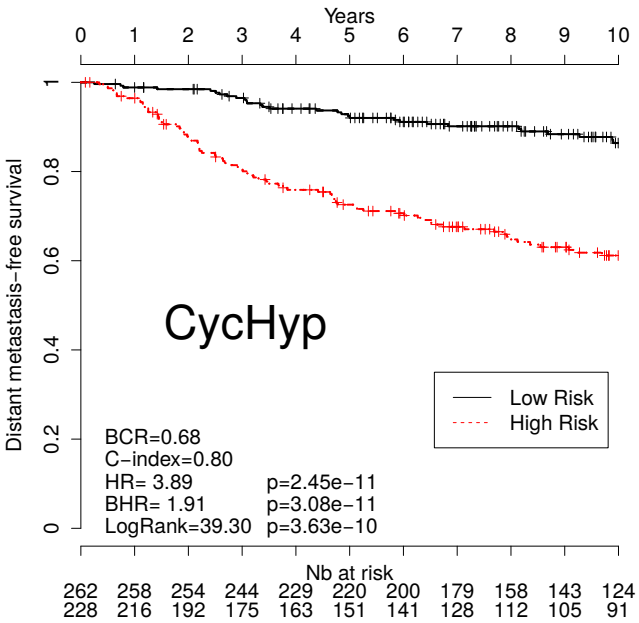


Figure A.7: Kaplan-Meier survival curves of patients with primary breast cancer, as determined by using the CycHyp signature. The validation set is reduced here to node-negative and untreated patients.

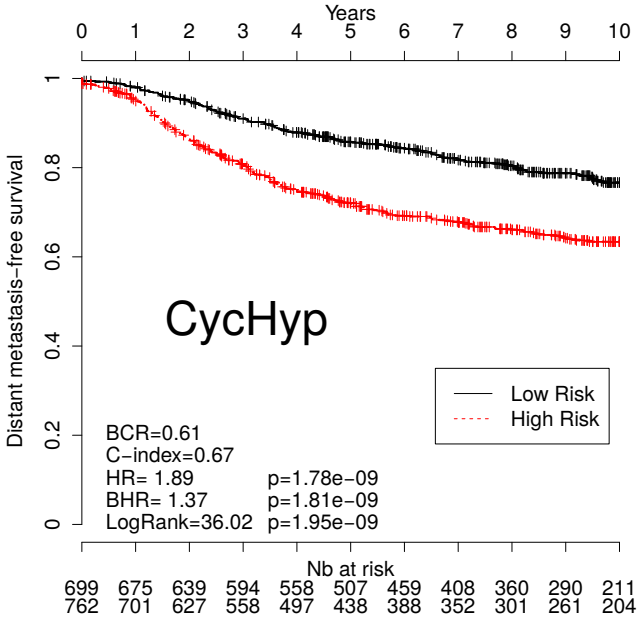
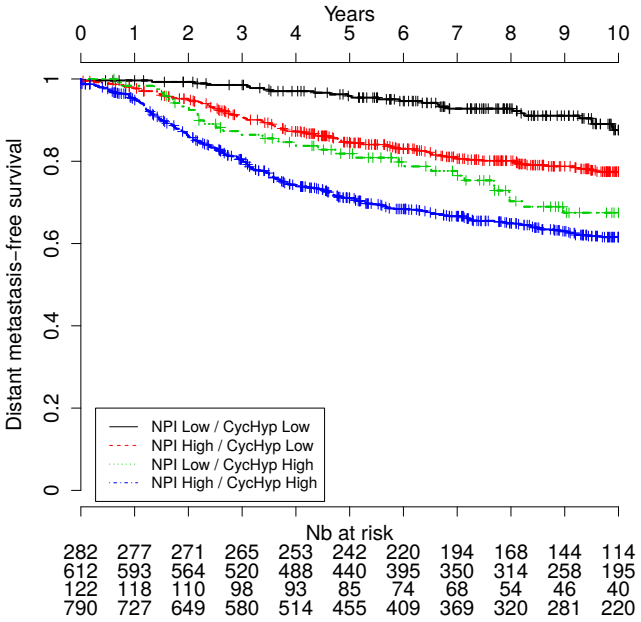
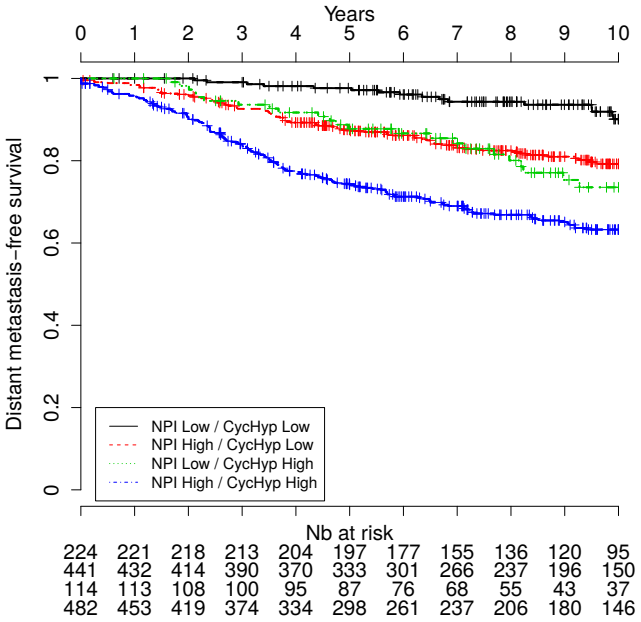


Figure A.8: Kaplan-Meier survival curves of patients with primary breast cancer, as determined by using the CycHyp signature. The validation set is defined here as all patients minus the best population of ER+/HER2-, node-negative and untreated patients.

A.9 The CycHyp signature in association with NPI

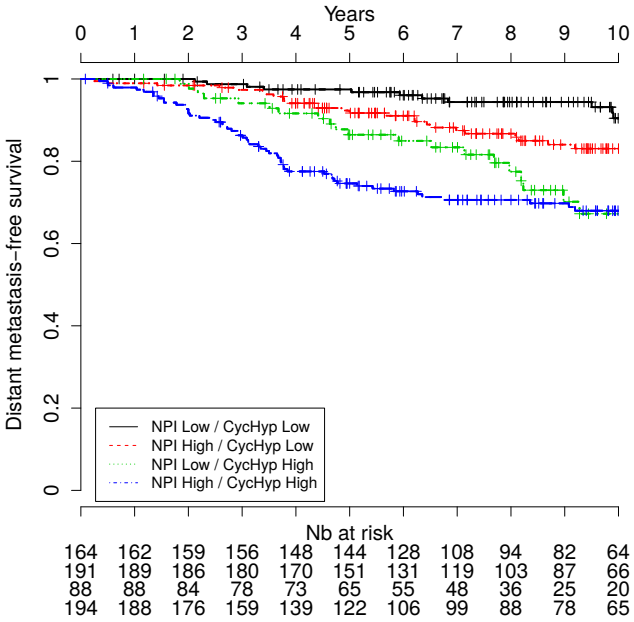


(a) All patients

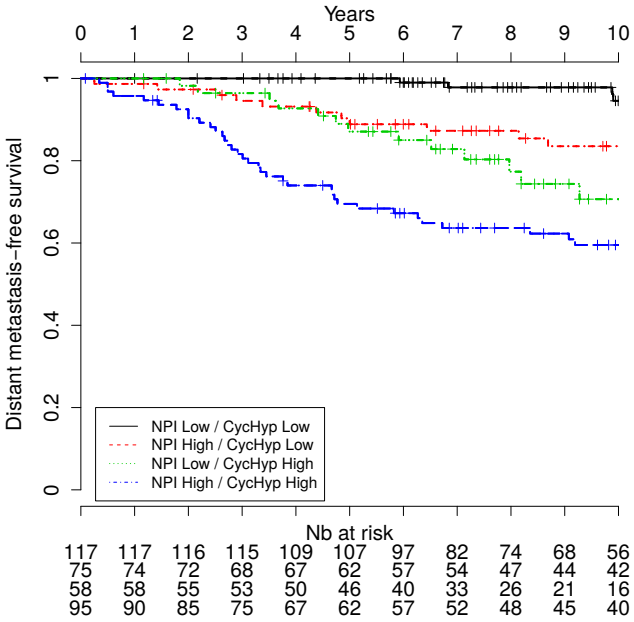


(b) ER+/HER2-

Figure A.9: Kaplan-Meier survival curves of patients with primary breast cancer stratified at low or high risk according to the CycHyp signature and the NPI nomenclature.

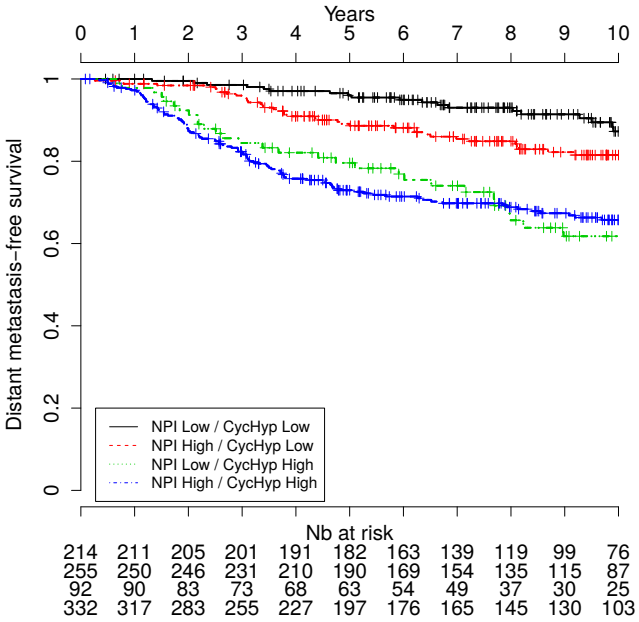


(a) ER+/HER2-, n-

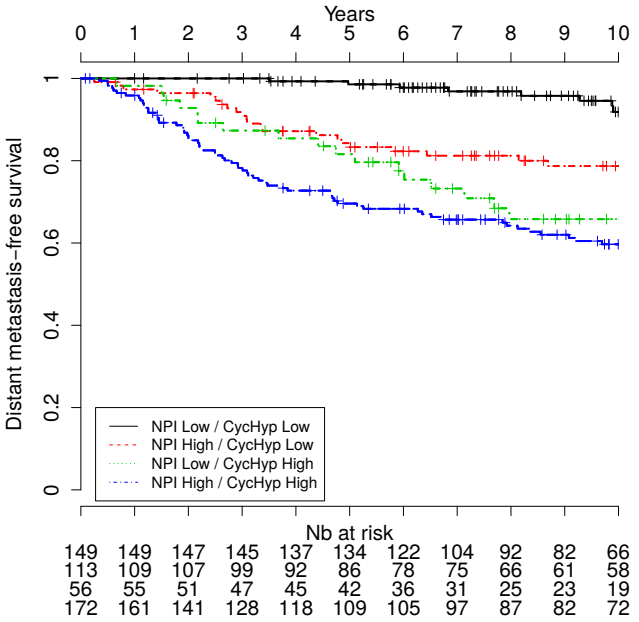


(b) ER+/HER2-, n-, untreated

Figure A.10: Kaplan-Meier survival curves of patients with primary breast cancer stratified at low or high risk according to the CycHyp signature and the NPI nomenclature.



(a) n-



(b) n-, untreated

Figure A.11: Kaplan-Meier survival curves of patients with primary breast cancer stratified at low or high risk according to the CycHyp signature and the NPI nomenclature.

A.10 The CycHyp signature in association with NPI with 6 groups

Blamey *et al.* [13] proposed to divided the Nottingham Prognostic Index (NPI) in six groups:

- Excellent Prognostic Group:

$$NPI \leq 2.4$$

- Good Prognostic Group:

$$2.4 < NPI \leq 3.4$$

- Moderate Prognostic Group 1:

$$3.4 < NPI \leq 4.4$$

- Moderate Prognostic Group 2:

$$4.4 < NPI \leq 5.4$$

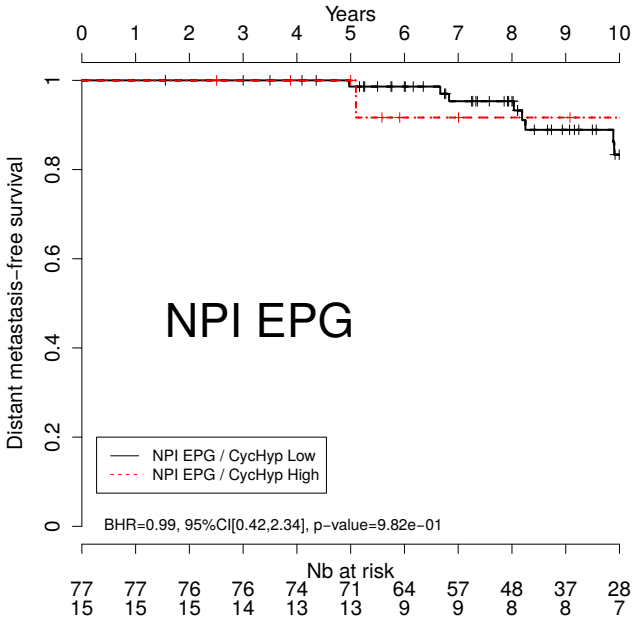
- Poor Prognostic Group:

$$5.4 < NPI \leq 6.4$$

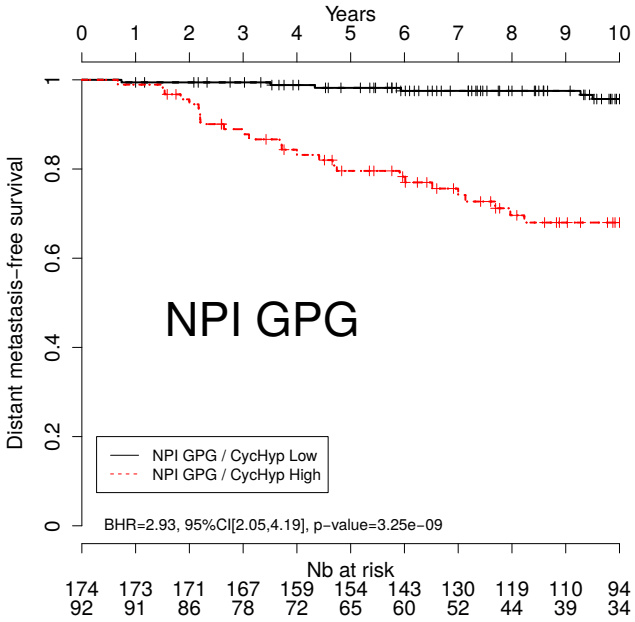
- Very poor Prognostic Group:

$$6.4 < NPI$$

We report in figures A.12, A.13 and A.14 the performances of the CycHyp signature in each of these six NPI-subgroups on the 8 validation data sets. The performances of the CycHyp signature are difficult to interpret in the 1st, 5th and 6th groups due to the small numbers of events. Very interesting results are obtained in the good prognostic group (GPG) with a BHR of 2.93 (p-values=3.25e-9). In this group, the CycHyp model is able to identify almost every patients with distant metastasis. Good performances are also obtained in moderate prognostic group 1 (MPG1) with a BHR of 1.61 (p-values=4.21e-5).

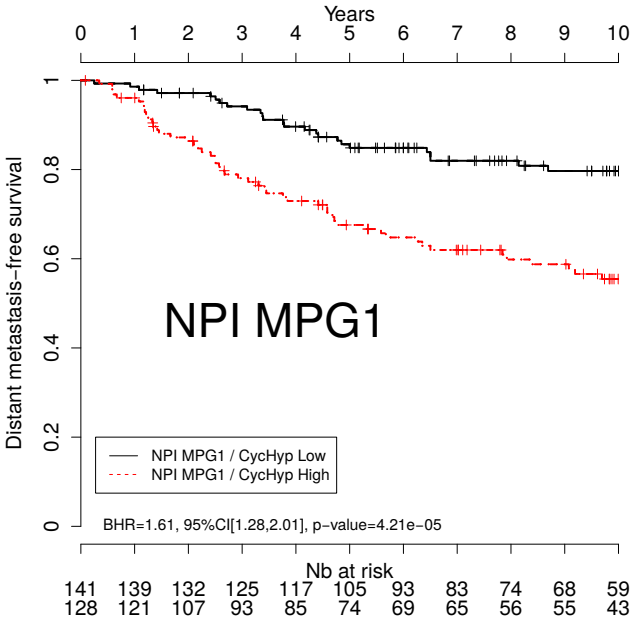


(a) Excellent Prognostic Group

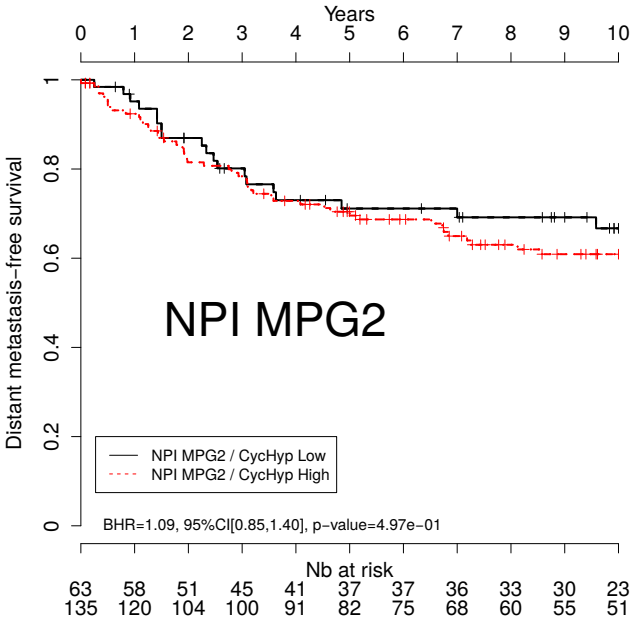


(b) Good Prognostic Group

Figure A.12: Kaplan-Meier survival curves of patients with primary breast cancer stratified at low or high risk according to the CycHyp signature and the NPI nomenclature.

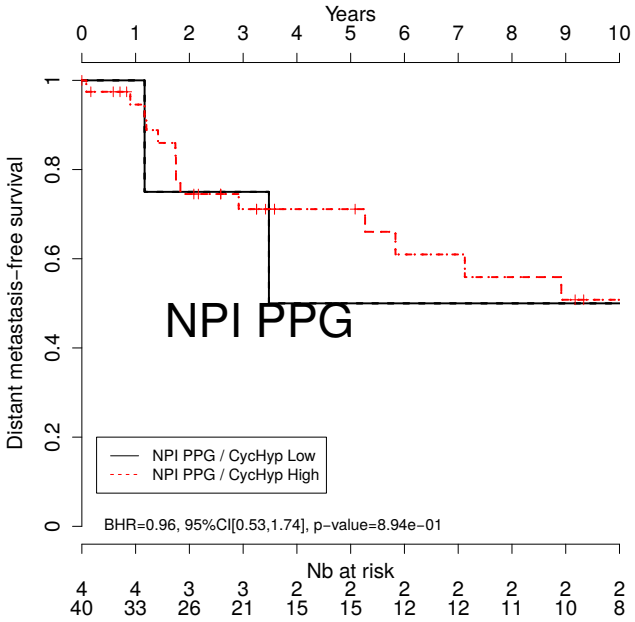


(a) Moderate Prognostic Group 1

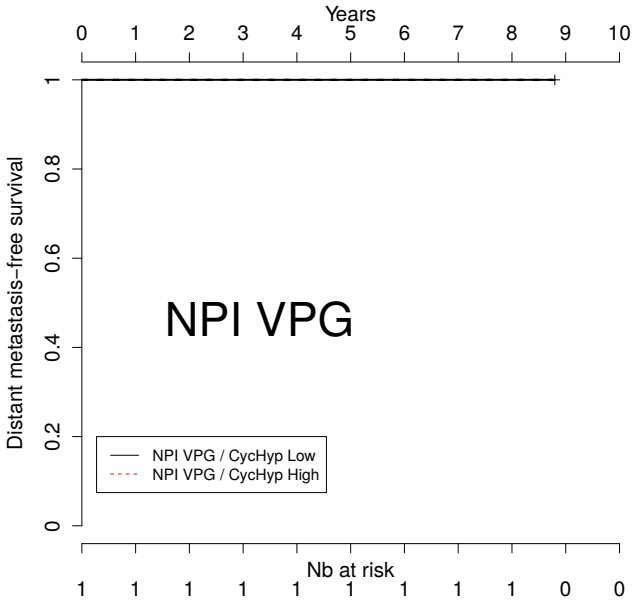


(b) Moderate Prognostic Group 2

Figure A.13: Kaplan-Meier survival curves of patients with primary breast cancer stratified at low or high risk according to the CycHyp signature and the NPI nomenclature.



(a) Poor Prognostic Group



(b) Very poor Prognostic Group

Figure A.14: Kaplan-Meier survival curves of patients with primary breast cancer stratified at low or high risk according to the CycHyp signature and the NPI nomenclature.

Chapter B

Supplementary Coxlogit

In this appendix, we describe the optimization procedure used in the estimation of the regularized Coxlogit model. First, we present the derivatives of the log-likelihoods of the Cox, logistic and Coxlogit models in respectively sections B.1, B.2 and B.3. Those derivatives are essential in the optimization algorithm presented in section B.4.

B.1 Cox partial log-likelihood and derivatives

For the sake of simplicity, we define two ensembles $R(t_i)$ and $C(t_i)$ that are used in the computation of the Cox partial log-likelihood:

$$R(t_i) = \{j | t_j \geq t_i\} \tag{B.1}$$
$$C(t_i) = \{j | \delta_j = 1, t_j \leq t_i\} \tag{B.2}$$

The partial log-likelihood presented here is similar to the one presented in section 3.3.1. However, it is here written with respect to the risk scores $\boldsymbol{z}$.

$$l_{\text{cox}}(\boldsymbol{z}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{z}_i - \log \sum_{k \in R(t_i)} \exp(\boldsymbol{z}_k) \right] \tag{B.3}$$

$$l'_{\text{cox}}(\boldsymbol{z})_j = \delta_j - \sum_{i \in C(t_j)} \frac{\exp(\boldsymbol{z}_j)}{\sum_{k \in R(t_i)} \exp(\boldsymbol{z}_k)} \tag{B.4}$$

$$l''_{\text{cox}}(\boldsymbol{z})_{jj} = - \sum_{i \in C(t_j)} \frac{\exp(\boldsymbol{z}_j) \sum_{k \in R(t_i)} \exp(\boldsymbol{z}_k) - (\exp(\boldsymbol{z}_j))^2}{\left(\sum_{k \in R(t_i)} \exp(\boldsymbol{z}_k) \right)^2} \tag{B.5}$$

B.2 Logistic model log-likelihood and derivatives

Similarly to the Cox model, we give here the logistic log-likelihood and its derivatives with respect to the risk scores $\boldsymbol{z}$:

$$l_{\text{logi}}(\boldsymbol{z}) = \sum_{i=1}^{n} \log(1 + \exp(-g_i z_i)) \tag{B.6}$$

$$l'_{\text{logi}}(\boldsymbol{z})_j = \frac{-g_i \exp(-g_i z_i)}{1 + \exp(-g_i z_i)} \tag{B.7}$$

$$l''_{\text{logi}}(\boldsymbol{z})_{jj} = \frac{\exp(-g_i z_i)}{(1 + \exp(-g_i z_i))^2} \tag{B.8}$$

B.3 Coxlogit model log-likelihood and derivatives

As presented in section 6.2.1, the Coxlogit model can be described as a mixture of the Cox and logistic model. The log-likelihood and its derivatives are thus also a mixture of the Cox and logistic ones presented in the previous sections.

$$l(\boldsymbol{z}) = (1 - \gamma)\ l_{\text{cox}}(\boldsymbol{z}) + \ \gamma\ l_{\text{logi}}(\boldsymbol{z}) \tag{B.9}$$

$$l'(\boldsymbol{z})_j = (1 - \gamma)\ l'_{\text{cox}}(\boldsymbol{z})_j + \ \gamma\ l'_{\text{logi}}(\boldsymbol{z})_j \tag{B.10}$$

$$l''(\boldsymbol{z})_{jj} = (1 - \gamma)\ l''_{\text{cox}}(\boldsymbol{z})_{jj} + \ \gamma\ l''_{\text{logi}}(\boldsymbol{z})_{jj} \tag{B.11}$$

B.4 Regularization path for generalized linear models

The Coxlogit model can be seen as generalized linear model similarly as the Cox and logistic regression. In this thesis, we implemented in R the algorithm for estimation of generalized linear models described in [52, 117] to fit the model parameters represented here by a vector λ.

The general principle of the algorithm B.1 is to follow the regularization path. The algorithm starts with a very high regularization constant λ such that the optimal solution is trivial ($\boldsymbol{\beta} = \mathbf{0}^{n \times 1}$). The λ constant is then slowly decreased to allows more and more non-zeros β_j (the model parameters).

Algorithm B.1: Iteratively reweighted least squares algorithm

Data: The normalized data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$.
Result: A sparse solution $\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \frac{1}{n} l(\boldsymbol{\beta}) - \lambda \mathcal{R}(\boldsymbol{\beta})$.
Initialize $\boldsymbol{\beta} = \mathbf{0}^{n \times 1}$ and set $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$;
Initialize λ sufficiently high such that the trivial solution is optimal;
while the number of non-zero $\hat{\beta}_j <$ **threshold do**
  Decrease λ;
  repeat
    Compute $l'(\boldsymbol{\eta})$ and $l''(\boldsymbol{\eta})$;
    $\boldsymbol{w} = \text{diag}(l''(\boldsymbol{\eta}))$;
    $\boldsymbol{z} = \boldsymbol{\eta} - \boldsymbol{w}^{-1} l'(\boldsymbol{\eta})$;
    With a coordinate descent algorithm, find $\hat{\boldsymbol{\beta}}$ minimizing:

$$\frac{1}{n} \sum_{i=1}^{n} w_i (z_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \mathcal{R}(\boldsymbol{\beta}) \tag{B.12}$$

    Set $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\eta} = \mathbf{X}\hat{\boldsymbol{\beta}}$;
  until convergence of $\hat{\boldsymbol{\beta}}$;
end

For each λ along the regularization path, the algorithm performs an iteratively reweighted least squares (IRLS). In these IRLS, the previous solution is used as warm start for the current λ. Exploiting these warm start allows us to have very few iterations for each step in the regularization path.

In the algorithm B.1, the optimization problem is thus reduced to repeatedly solving a penalized least-square:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} w_i (z_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \mathcal{R}(\boldsymbol{\beta}) \tag{B.13}$$

This penalized least-square can be efficiently solved with a coordinate descent, similarly as what is used in [52, 117]. This coordinate descent algorithm was first proposed by [139] for penalized linear regression.

Chapter C

Supplementary results with the balanced hazard ratio

This section reports additional results and comparisons to chapter 8. In particular we extend figure 8.7 while comparing the hazard ratio, D-index [108], SEP [112] and the BHR [17]. We decide to represent the BHR and SEP to the square to have the four metrics on the same scale. With respect to the hazard ratio, the D-index slightly corrects the performance with unbalanced risk groups. This correction is much stronger with the BHR and SEP. These two metrics tend to zero when all patients are in the same risk group. Figure C.2 compares the D-index and BHR^2 with their 95% confidence intervals. The confidence intervals of the D-index is much wider than the BHR^2 one.

To further compare the four metrics, we can repeat the experiment presented is section 8.8. In these experiments, we generate survival data for which an underlying threshold between risk groups is **a priori** fixed according to a prescribed proportion between a high or low risk profile.

Figure C.3 offers a closer look at the distribution over 500 runs of the proportions between risk groups for which respectively the HR, D-index, BHR and SEP, is maximum. The true proportion ρ in the low risk group was here fixed to 80%. The maximal BHR is clearly more concentrated around the true underlying proportion while the maximum HR distribution is much more dispersed and skewed towards an excessively large value. The maximum D-index distribution is very similar to what is observed with the hazard ratio. At the opposite, the maximum SEP distribution is skewed toward a proportion of 0.5. These results tend to show that the balanced hazard ratio is better at finding the true proportion between risk groups.

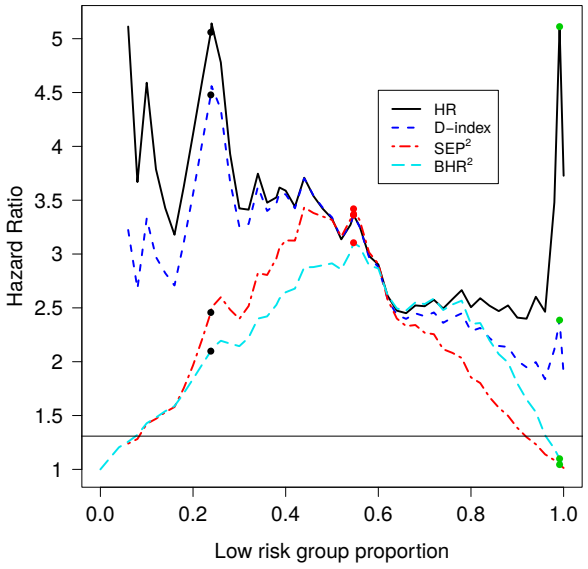


Figure C.1: Performance metrics on the VDX dataset while varying the proportions in each risk group.

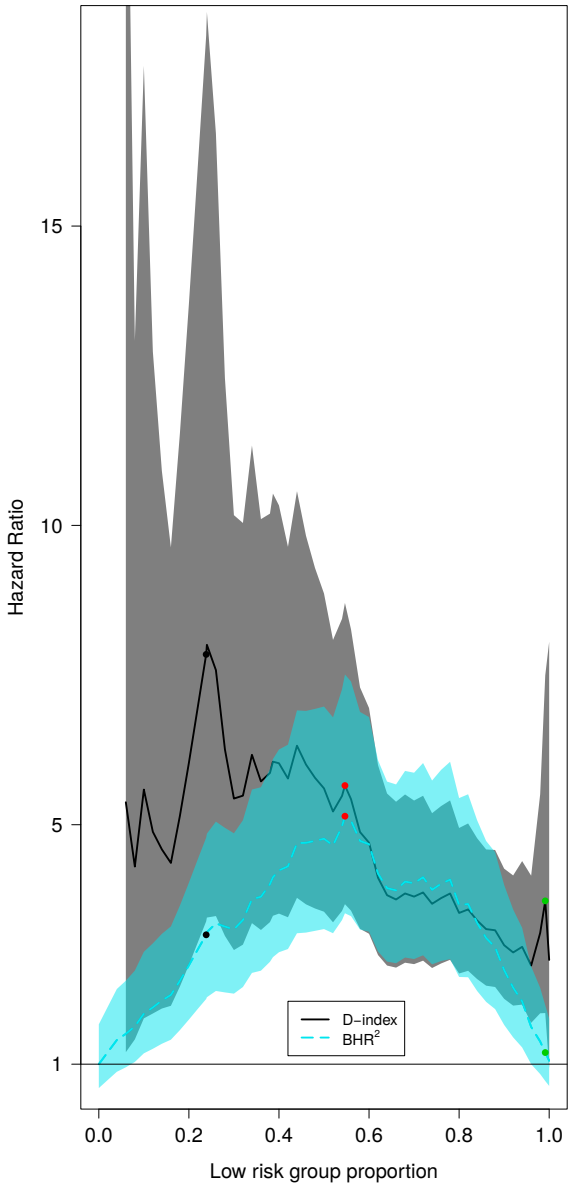
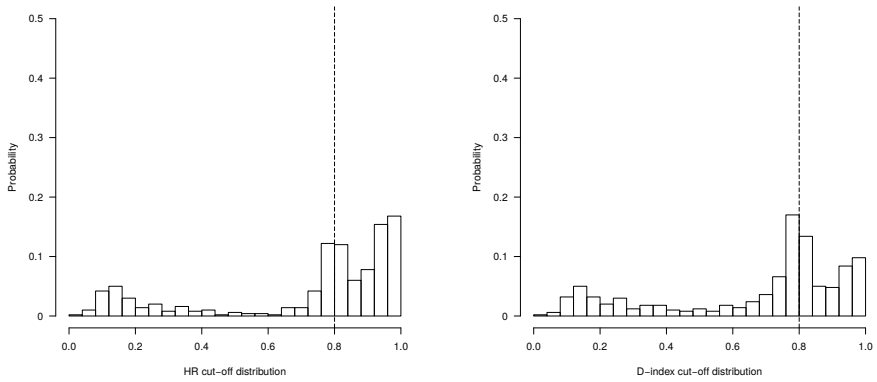
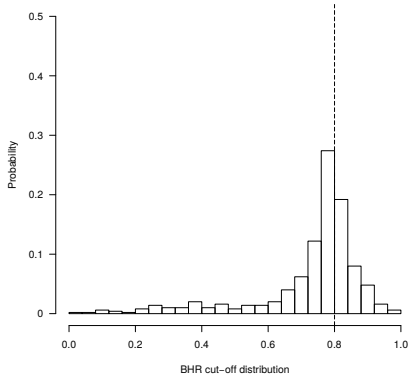


Figure C.2: D-index, squared balanced hazard ratio and their confidence intervals on the VDX dataset while varying the proportions in each risk group. The balanced hazard ratio is squared to be on the same scale as the hazard ratio and the D-index.

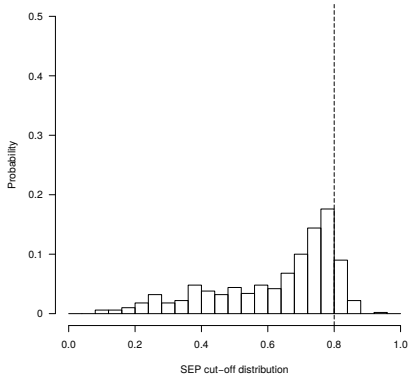


(a) Hazard ratio

(b) D-index



(c) Balanced hazard ratio



(d) SEP

Figure C.3: Distribution over 500 runs of the low risk group proportion for which respectively the BHR (c), HR (a), SEP (d) and D-index (b), is maximum. The experiments are conducted with $n = 200$ patients, the shape parameter $k = 1$ and the true group hazard ratio $\exp(\mu) = 3$. The true underlying proportion of patients in the low risk group was set to 80%.

References

[1] Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics (Oxford, England)*, **26**(3), 392–8.

[2] Altman, D. G. (1991). Categorising continuous variables. *British journal of cancer*, **64**(5), 975.

[3] Altman, D. G., McShane, L. M., Sauerbrei, W., and Taube, S. E. (2012). Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Medicine*, **10**(1), 51.

[4] Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(10), 6562–6566.

[5] Baudelet, C., Ansiaux, R., Jordan, B. F., Havaux, X., Macq, B., and Gallez, B. (2004). Physiological noise in murine solid tumours using T2*-weighted gradient-echo imaging: a marker of tumour acute hypoxia? *Physics in medicine and biology*, **49**(15), 3389–3411.

[6] Baudelet, C., Cron, G. O., Ansiaux, R., Crokart, N., DeWever, J., Feron, O., and Gallez, B. (2006). The role of vessel maturation and vessel functionality in spontaneous fluctuations of T2*-weighted GRE signal within tumors. *NMR in Biomedicine*, **19**(1), 69–76.

[7] Beck, A. H., Knoblauch, N. W., Hefti, M. M., Kaplan, J., Schnitt, S. J., Culhane, A. C., Schroeder, M. S., Risch, T., Quackenbush, J., and Haibe-Kains, B. (2013). Significance analysis of prognostic signatures. *PLoS computational biology*, **9**(1), e1002875.

[8] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statisitical Society, Series B*, **57**(1), 289–300.

[9] Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal ofMachine Learning Research*, **13**, 281–305.

[10] Bertout, J. a., Patel, S. a., and Simon, M. C. (2008). The impact of O2 availability on human cancer. *Nature reviews. Cancer*, **8**(12), 967–975.

[11] Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, **17**(10), 1169–1186.

[12] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, volume 4.

[13] Blamey, R. W., Ellis, I. O., Pinder, S. E., Lee, a. H. S., Macmillan, R. D., Morgan, D. a. L., Robertson, J. F. R., Mitchell, M. J., Ball, G. R., Haybittle, J. L., and Elston, C. W. (2007). Survival of invasive breast cancer according to the Nottingham Prognostic Index in cases diagnosed in 1990-1999. *European Journal of Cancer*, **43**(10), 1548–1555.

[14] Boidot, R., Vegran, F., Meulle, A., Le Breton, A., Dessy, C., Sonveaux, P., Lizard-Nacol, S., and Feron, O. (2012). Regulation of monocarboxylate transporter MCT1 expression by p53 mediates inward and outward lactate fluxes in tumors. *Cancer Research*, **72**(4), 939–948.

[15] Boidot, R., Branders, S., Helleputte, T., Rubio, L. I., Dupont, P., and Feron, O. (2014). A generic cycling hypoxia-derived prognostic gene signature: application to breast cancer profiling. *Oncotarget*, **5**(16).

[16] Boulesteix, A. L., Guillemot, V., and Sauerbrei, W. (2011). Use of pretransformation to cope with extreme values in important candidate features. *Biometrical Journal*, **53**(4), 673–688.

[17] Branders, S. and Dupont, P. (2015). A balanced hazard ratio for risk group evaluation from survival data. *Statistics in Medicine*, **34**(17).

[18] Branders, S., D'Ambrosio, R., and Dupont, P. (2014). The Coxlogit model: Feature selection from survival and classification data. In *2014*

IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making (MCDM), pages 137–143.

[19] Branders, S., D'Ambrosio, R., and Dupont, P. (2015a). A mixture Cox-Logistic model for feature selection from survival and classification data. *arXiv:1502.01493 [stat.ML]*, pages 1–6.

[20] Branders, S., Frénay, B., and Dupont, P. (2015b). Survival Analysis with Cox Regression and Random Non-linear Projections. *Proceedings of the 23th European Symposium on Artificial Neural Networks*, pages 119–124.

[21] Breslow, N. E. (1972). Contribution to the discussion of the paper by DR Cox. *Journal of the Royal Statistical Society. Series B*, **34**(2), 216–217.

[22] Bristow, R. G. and Hill, R. P. (2008). Hypoxia and metabolism: hypoxia, DNA repair and genetic instability. *Nature reviews. Cancer*, **8**(3), 180–192.

[23] Buffa, F. M., Harris, a. L., West, C. M., and Miller, C. J. (2010). Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *British journal of cancer*, **102**(2), 428–435.

[24] Cañette, I. (2009). Cox model and conditional logistic model, back and forth (part 1). url: https://stayconsistent.wordpress.com /2009/01/25/cox-model-and-conditional-logistic-model-back-and-forth/. Accessed: 2015-08-30.

[25] Chi, J.-T., Wang, Z., Nuyten, D. S. a., Rodriguez, E. H., Schaner, M. E., Salim, A., Wang, Y., Kristensen, G. B., Helland, A., Bø rresen Dale, A.-L., Giaccia, A., Longaker, M. T., Hastie, T., Yang, G. P., van de Vijver, M. J., and Brown, P. O. (2006). Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS medicine*, **3**(3), e47.

[26] Chitneni, S. K., Palmer, G. M., Zalutsky, M. R., and Dewhirst, M. W. (2011). Molecular imaging of hypoxia. *Journal of Nuclear Medicine*, **52**(2), 165–168.

[27] Collett, D. (2003a). *Modelling binary data*. CRC press.

[28] Collett, D. (2003b). *Modelling Survival Data in Medical Research*. Chapman&Hall-CRC.

[29] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297.

[30] Cox, D. (1972). Regression models and life tables. *J R Stat Soc B*, **34**(2), 187–220.

[31] Cox, D. R. (1975). Partial Likelihood. *Biometrika*, **62**, 269–276.

[32] Cox, D. R. and Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

[33] Daneau, G., Boidot, R., Martinive, P., and Feron, O. (2010). Identification of cyclooxygenase-2 as a major actor of the transcriptomic adaptation of endothelial and tumor cells to cyclic hypoxia: Effect on angiogenesis and metastases. *Clinical Cancer Research*, **16**(2), 410–419.

[34] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, **7**, 1–30.

[35] Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., D'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A. L., Klijn, J. G. M., Foekens, J. a., Cardoso, F., Piccart, M. J., Buyse, M., and Sotiriou, C. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **13**(11), 3207–14.

[36] Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research*, **14**(16), 5158–5165.

[37] Dewhirst, M. W. (2007). Intermittent hypoxia furthers the rationale for hypoxia-inducible factor-1 targeting. *Cancer Research*, **67**(3), 854–855.

[38] Dewhirst, M. W. (2009). Relationships between Cycling Hypoxia, HIF-1, Angiogenesis and Oxidative Stress. *Radiation research*, **172**(6), 653–665.

[39] Dewhirst, M. W., Cao, Y., and Moeller, B. (2008). Cycling hypoxia and free radicals regulate angiogenesis and radiotherapy response. *Nature reviews. Cancer*, **8**(6), 425–437.

[40] Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, **72**, 557–565.

[41] Eng-Wong, J. and Isaacs, C. (2010). Prediction of benefit from adjuvant treatment in patients with breast cancer. *Clinical breast cancer*, **10 Suppl 1**, E32–E37.

[42] Espinosa, E., Vara, J. A. F., Navarro, I. S., Gámez-Pozo, A., Pinto, A., Zamora, P., Redondo, A., and Feliu, J. (2011). Gene profiling in breast cancer: Time to move forward.

[43] Essaghir, A. and Demoulin, J.-B. (2012). A minimal connected network of transcription factors regulated in human tumors and its application to the quest for universal cancer biomarkers. *PloS one*, **7**(6), e39666.

[44] Evers, L. and Messow, C. M. (2008). Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, **24**, 1632–1638.

[45] Favaro, E., Lord, S., Harris, A. L., and Buffa, F. M. (2011). Gene expression and hypoxia in breast cancer. *Genome Medicine*, **3**(8), 55.

[46] Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W. W., Comber, H., Forman, D., and Bray, F. (2013). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *European Journal of Cancer*, **49**(6), 1374–1403.

[47] Feron, O. (2009). Pyruvate into lactate and back: From the Warburg effect to symbiotic energy fuel exchange in cancer cells.

[48] Feron, O., Boidot, R., Branders, S., Dupont, P., and Helleputte, T. (2015). Signature of cycling hypoxia and use thereof for the prognosis of cancer. WO Patent App. PCT/EP2014/066,643.

[49] Fine, J. P. and Gray, R. J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, **94**(446), 496–509.

[50] Frénay, B. and Verleysen, M. (2010). Using svms with randomised feature spaces: an extreme learning approach. In *Proceedings of The 18th European Symposium on Artificial Neural Networks (ESANN)*, pages 315–320.

[51] Frénay, B. and Verleysen, M. (2011). Parameter-insensitive kernel in extreme learning for non-linear support vector regression. *Neurocomputing*, **74**(16), 2526 – 2531.

[52] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**(1), 1.

[53] Fu, W. J. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, **7**(3), 397–416.

[54] Galea, M. H., Blamey, R. W., Elston, C. E., and Ellis, I. O. (1992). The Nottingham Prognostic Index in primary breast cancer. *Breast cancer research and treatment*, **22**(3), 207–19.

[55] Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical journal. Biometrische Zeitschrift*, **52**(1), 70–84.

[56] Grandjean, M., Sermeus, A., Branders, S., Defresne, F., Dieu, M., Dupont, P., Raes, M., De Ridder, M., and Feron, O. (2013). Hypoxia Integration in the Serological Proteome Analysis Unmasks Tumor Antigens and Fosters the Identification of Anti-Phospho-eEF2 Antibodies as Potential Cancer Biomarkers. *PloS one*, **8**(10), e76508.

[57] Green, P. J. and Yandell, B. S. (1985). Semi-parametric Generalized Linear Models. pages 44–55.

[58] Guyon, I. (2006). *Feature Extraction: Foundations and Applications*, volume 207. Springer.

[59] Haibe-Kains, B. (2009). Identification and Assessment of Gene Signatures in Human Breast Cancer. *Phd thesis*.

[60] Haibe-Kains, B., Desmedt, C., Sotiriou, C., and Bontempi, G. (2008). A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics (Oxford, England)*, **24**(19), 2200–8.

[61] Haibe-Kains, B., Desmedt, C., Rothé, F., Piccart, M., Sotiriou, C., and Bontempi, G. (2010). A fuzzy gene expression-based computational approach improves breast cancer prognostication. *Genome biology*, **11**(2), R18.

[62] Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, **15**, 361–387.

[63] Hemingway, H., Croft, P., Perel, P., Hayden, J. a., Abrams, K., Timmis, A., Briggs, A., Udumyan, R., Moons, K. G. M., Steyerberg, E. W., Roberts, I., Schroter, S., Altman, D. G., Riley, R. D., and for the Progress Group (2013). Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *Bmj*, **5595**(February 2013), forthcoming.

[64] Hingorani, A. D. A., Windt, D. V. D. a. V. D., Riley, R. D., Abrams, K., Moons, K. G. M., Steyerberg, E. W., Schroter, S., Sauerbrei, W., Altman, D. G., and Hemingway, H. (2013). Prognosis research strategy (PROGRESS) 4: stratified medicine research. *Bmj*, **5793**(February 2013), 1–9.

[65] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55–67.

[66] Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, **16**(9), 965–980.

[67] Huang, D. W., Sherman, B. T., and Lempicki, R. a. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**(1), 1–13.

[68] Huang, G., Wang, D., and Lan, Y. (2011). Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, **2**(2), 107–122.

[69] Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, **70**(1-3), 489–501.

[70] Ignatiadis, M., Singhal, S. K., Desmedt, C., Haibe-Kains, B., Criscitiello, C., Andre, F., Loi, S., Piccart, M., Michiels, S., and Sotiriou, C. (2012). Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: A pooled analysis. *Journal of Clinical Oncology*, **30**(16), 1996–2004.

[71] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, **2**(3), 841–860.

[72] Japkowicz, N. and Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

[73] Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., and Guedj, M. (2010). Should we abandon the t-Test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE*, **5**(9), 1–9.

[74] Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC bioinformatics*, **7**, 359.

[75] Justice, A. C., Covinsky, K. E., and Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, **130**(6), 515–524.

[76] Kao, K.-J., Chang, K.-M., Hsu, H.-C., and Huang, A. T. (2011). Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC cancer*, **11**(1), 143.

[77] Kaplan, E. L. and Meyer, P. (1958). Nonparametric estimation from incomplete samples. *Journal of the American statistical association*, **53**(282), 457–481.

[78] Kohavi, R. and Sommerfield, D. (1995). Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. *First International Conference on Knowledge Discovery and Data Mining*, pages 192–197.

[79] Koppenol, W. H., Bounds, P. L., and Dang, C. V. (2011). Otto Warburg's contributions to current concepts of cancer metabolism. *Nature reviews. Cancer*, **11**(5), 325–337.

[80] Krishna, M. C., Matsumoto, S., Yasui, H., Saito, K., Devasahayam, N., Subramanian, S., and Mitchell, J. B. (2012). Electron paramagnetic resonance imaging of tumor pO2. *Radiation research*, **177**(4), 376–386.

[81] Leontieva, O. V. and Blagosklonny, M. V. (2012). Hypoxia and gerosuppression: The mTOR saga continues. *Cell Cycle*, **11**(21), 3926–3931.

[82] Leontieva, O. V., Natarajan, V., Demidenko, Z. N., Burdelya, L. G., Gudkov, a. V., and Blagosklonny, M. V. (2012). Hypoxia suppresses conversion from proliferative arrest to cellular senescence. *Proceedings of the National Academy of Sciences*, **109**(33), 13314–13318.

[83] Levine, A. J. (1997). P53, the Cellular Gatekeeper for Growth and Division. *Cell*, **88**(3), 323–331.

[84] Li, H. and Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 65–76.

[85] Liu, J. C., Egan, S. E., and Zacksenhaus, E. (2013). A Tumor initiating cell-enriched prognostic signature for HER2+:ERα- breast cancer; rationale, new features, controversies and future directions. *Oncotarget*, **4**(8), 1317–1328.

[86] Liu, Q., He, Q., and Shi, Z. (2008). Extreme support vector machine classifier. In *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining*, pages 222–233, Berlin, Heidelberg.

[87] Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J. a., Klijn, J. G. M., Larsimont, D., Buyse, M., Bontempi, G., Delorenzi, M., Piccart, M. J., and Sotiriou, C. (2007). Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **25**(10), 1239–46.

[88] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719–748.

[89] Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J.-F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P., and Boige, V. (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine*, **10**(5), e1001453.

[90] Martinive, P., Defresne, F., Bouzin, C., Saliez, J., Lair, F., Grégoire, V., Michiels, C., Dessy, C., and Feron, O. (2006a). Preconditioning of

the tumor vasculature and tumor cells by intermittent hypoxia: Implications for anticancer therapies. *Cancer Research*, **66**(24), 11736–11744.

[91] Martinive, P., De Wever, J., Bouzin, C., Baudelet, C., Sonveaux, P., Grégoire, V., Gallez, B., and Feron, O. (2006b). Reversal of temporal and spatial heterogeneities in tumor perfusion identifies the tumor vascular tone as a tunable variable to improve drug delivery. *Molecular cancer therapeutics*, **5**(6), 1620–1627.

[92] Micel, L. N., Tentler, J. J., Smith, P. G., and Eckhardt, S. G. (2013). Role of ubiquitin ligases and the proteasome in oncogenesis: Novel targets for anticancer therapies. *Journal of Clinical Oncology*, **31**(9), 1231–1238.

[93] Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C., and Lendasse, A. (2009). Op-elm: Optimally pruned extreme learning machine. *Neural Networks, IEEE Transactions on*, **21**(1), 158–162.

[94] Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., and Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(38), 13550–5.

[95] Minn, A. J., Gupta, G. P., Padua, D., Bos, P., Nguyen, D. X., Nuyten, D., Kreike, B., Zhang, Y., Wang, Y., Ishwaran, H., Foekens, J. a., van de Vijver, M., and Massagué, J. (2007). Lung metastasis genes couple breast tumor size and metastatic spread. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(16), 6740–5.

[96] Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrå le, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, **34**(3), 267–273.

[97] Ng, A. Y. (1998). On Feature selection: Learning with Exponentially many Irrelevant Features Training Examples. *Proc. 15th Interntional Conference on Machine Learning*, pages 404–412.

[98] Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., and Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine*, **351**(27), 2817–26.

[99] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, **22**(10), 1345–1359.

[100] Prat, A., Ellis, M. J., and Perou, C. M. (2011). Practical implications of gene-expression-based assays for breast oncologists.

[101] Prentice, R. and Kalbfleisch, J. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34**(4), 541–554.

[102] Rakha, E. a., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V., Fox, S. B., Ichihara, S., Jacquemier, J., Lakhani, S. R., Palacios, J., Richardson, A. L., Schnitt, S. J., Schmitt, F. C., Tan, P.-H., Tse, G. M., Badve, S., and Ellis, I. O. (2010). Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast cancer research : BCR*, **12**(4), 207.

[103] Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hewlett, J., Gerson, N., and Parker, H. L. (2001). Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology*, **19**(4), 980–991.

[104] Reilly, B. and Evans, A. (2006). Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of Internal Medicine*, **144**(3), 201.

[105] Reis-Filho, J. S. and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, **378**(9805), 1812–1823.

[106] Riley, R. D., Hayden, J. a., Steyerberg, E. W., Moons, K. G. M., Abrams, K., Kyzas, P. A., Malats, N., Briggs, A., Schroter, S., Altman, D. G., and Hemingway, H. (2013). Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Medicine*, **10**(2), e1001380.

[107] Rosset, S., Zhu, J., and Hastie, T. J. (2004). Margin Maximizing Loss Functions. *Advances in Neural Information Processing Systems 16*, pages 1237–1244.

[108] Royston, P. and Sauerbrei, W. (2004). A new measure of prognostic separation in survival data. *Statistics in medicine*, **23**(5), 723–48.

[109] Royston, P. and Sauerbrei, W. (2007). Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. *Computational Statistics & Data Analysis*, **51**(9), 4240–4253.

[110] Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, **25**(October 2005), 127–141.

[111] Sabatier, R., Finetti, P., Cervera, N., Lambaudie, E., Esterni, B., Mamessier, E., Tallet, A., Chabannon, C., Extra, J.-M., Jacquemier, J., Viens, P., Birnbaum, D., and Bertucci, F. (2011). A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast cancer research and treatment*, **126**(2), 407–20.

[112] Sauerbrei, W., Hübner, K., Schmoor, C., and Schumacher, M. (1997). Validation of existing and development of new prognostic classification schemes in node negative breast cancer. *Breast cancer research and treatment*, **42**(2), 149–163.

[113] Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J. G., Kölbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research*, **68**(13), 5405–13.

[114] Seigneuric, R., Starmans, M. H. W., Fung, G., Krishnapuram, B., Nuyten, D. S. a., van Erk, A., Magagnin, M. G., Rouschop, K. M., Krishnan, S., Rao, R. B., Evelo, C. T. a., Begg, A. C., Wouters, B. G., and Lambin, P. (2007). Impact of supervised gene signatures of early hypoxia on patient survival. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*, **83**(3), 374–82.

[115] Semenza, G. L. (2011). Oxygen Sensing, Homeostasis, and Disease. *New England Journal of Medicine*, **365**(6), 537–547.

[116] Seront, E., Rottey, S., Sautois, B., Kerger, J., D'Hondt, L. a., Verschaeve, V., Canon, J.-L., Dopchie, C., Vandenbulcke, J. M., Whenham, N., Goeminne, J. C., Clausse, M., Verhoeven, D., Glorieux, P., Branders, S., Dupont, P., Schoonjans, J., Feron, O., and Machiels, J.-P. (2012). Phase II study of everolimus in patients with locally advanced or metastatic transitional cell carcinoma of the urothelial

tract: clinical activity, molecular response, and biomarkers. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, **23**(10), 2663—-2670.

[117] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of statistical software*, **39**(5).

[118] Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, **3**(1), Article3.

[119] Snijders, A. M., Langley, S., Mao, J.-H., Bhatnagar, S., Bjornstad, K. a., Rosen, C. J., Lo, A., Huang, Y., Blakely, E. a., Karpen, G. H., Bissell, M. J., and Wyrobek, A. J. (2014). An interferon signature identified by RNA-sequencing of mammary tissues varies across the estrous cycle and is predictive of metastasis-free survival. *Oncotarget*, **5**(12), 4011–25.

[120] Snoek, J., Larochelle, H., and Adams, R. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Nips*, pages 1–9.

[121] Sonveaux, P., Végran, F., Schroeder, T., Wergin, M. C., Verrax, J., Rabbani, Z. N., De Saedeleer, C. J., Kennedy, K. M., Diepart, C., Jordan, B. F., Kelley, M. J., Gallez, B., Wahl, M. L., Feron, O., and Dewhirst, M. W. (2008). Targeting lactate-fueled respiration selectively kills hypoxic tumor cells in mice. *Journal of Clinical Investigation*, **118**(12), 3930–3942.

[122] Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer. *The New England journal of medicine*, **360**(8), 790–800.

[123] Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(18), 10393–10398.

[124] Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M., and Delorenzi, M. (2006).

Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**(4), 262–72.

[125] Sparano, J. A. and Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials.

[126] Starmans, M. H. W., Chu, K. C., Haider, S., Nguyen, F., Seigneuric, R., Magagnin, M. G., Koritzinsky, M., Kasprzyk, A., Boutros, P. C., Wouters, B. G., and Lambin, P. (2012). The prognostic value of temporal in vitro and in vivo derived hypoxia gene-expression signatures in breast cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*, **102**(3), 436–43.

[127] Steyerberg, E., Moons, K. G. M., van der Windt, D., Hayden, J., Perel, P., Schroter, S., Riley, R., Hemingway, H., and Altman, R. B. (2012). Prognosis research strategy (PROGRESS) series 3: prognostic models. *British Medical Journal*, **10**(2).

[128] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, **21**(1), 128–138.

[129] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545–15550.

[130] Symmans, W. F., Hatzis, C., Sotiriou, C., Andre, F., Peintinger, F., Regitnig, P., Daxenbichler, G., Desmedt, C., Domont, J., Marth, C., Delaloge, S., Bauernhofer, T., Valero, V., Booser, D. J., Hortobagyi, G. N., and Pusztai, L. (2010). Genomic index of sensitivity to endocrine therapy for breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **28**(27), 4111–9.

[131] Thornton, C., Hutter, F., Hoos, H. H., Leyton-Brown, K., and Chris Thornton, Frank Hutter, Holger H. Hoos, K. L.-B. (2013). Auto-WEKA: Combined Selection and Hyperparameter Optimization of

Classification Algorithms. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 847–855.

[132] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (...*, **58**(1), 267–288.

[133] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine*, **16**(4), 385–95.

[134] Troester, M. a., Hoadley, K. a., Sø rlie, T., Herbert, B. S., Bø rresen Dale, A. L., Lø nning, P. E., Shay, J. W., Kaufmann, W. K., and Perou, C. M. (2004). Cell-type-specific responses to chemotherapeutics in breast cancer. *Cancer Research*, **64**(12), 4218–4226.

[135] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(9), 5116–5121.

[136] Van Belle, V., Pelckmans, K., Suykens, J., and Van Huffel, S. (2007). Support Vector Machine for Survival Analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, pages 1–8.

[137] Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. a. K. (2011a). Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics (Oxford, England)*, **27**(1), 87–94.

[138] Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. a. K. (2011b). Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, **53**(2), 107–118.

[139] van der Kooij, A. J. (2007). *Prediction accuracy and stability of regression with optimal scaling transformations*. Ph.D. thesis.

[140] van Noorden, R., Maher, B., and Nuzzo, R. (2014). The top 100 papers. *Nature*, (4), 4–7.

[141] van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. a. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Lins-

ley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–6.

[142] Venet, D., Dumont, J. E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, **7**(10), e1002240.

[143] Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D., and Foekens, J. a. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**(9460), 671–9.

[144] Winter, S. C., Buffa, F. M., Silva, P., Miller, C., Valentine, H. R., Turley, H., Shah, K. A., Cox, G. J., Corbridge, R. J., Homer, J. J., Musgrove, B., Slevin, N., Sloan, P., Price, P., West, C. M. L., and Harris, A. L. (2007). Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Research*, **67**(7), 3441–3449.

[145] Wise, D. R., Ward, P. S., Shay, J. E. S., Cross, J. R., Gruber, J. J., Sachdeva, U. M., Platt, J. M., DeMatteo, R. G., Simon, M. C., and Thompson, C. B. (2011). Hypoxia promotes isocitrate dehydrogenase-dependent carboxylation of alpha-ketoglutarate to citrate to support cell growth and viability. *Proceedings of the National Academy of Sciences*, **108**(49), 19611–19616.

[146] Yasui, H., Matsumoto, S., Devasahayam, N., Munasinghe, J. P., Choudhuri, R., Saito, K., Subramanian, S., Mitchell, J. B., and Krishna, M. C. (2010). Low-field magnetic resonance imaging to visualize chronic and cycling hypoxia in tumor-bearing mice. *Cancer Research*, **70**(16), 6427–6436.

[147] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.